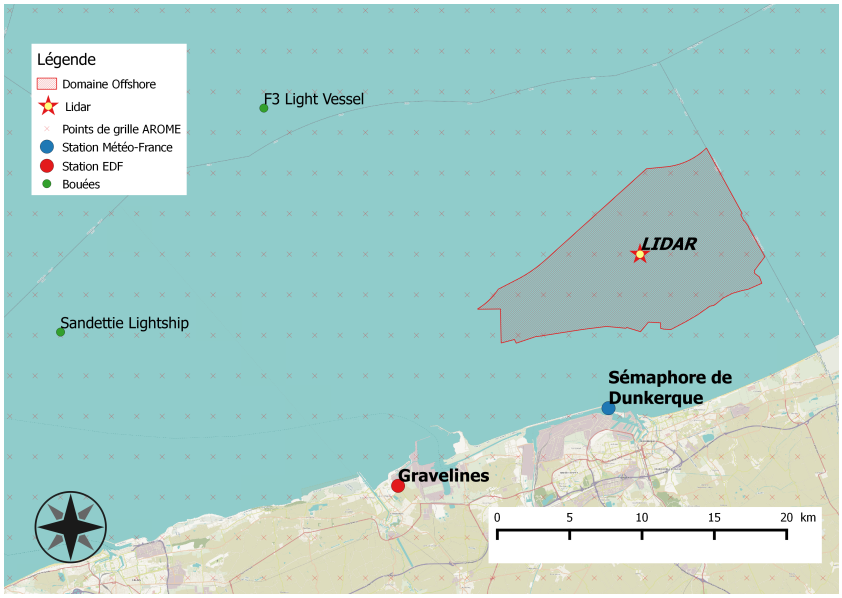
Projet de parc éolien off-shore au large de Dunkerque

Extension de la série d'observations horaires



Version 2 du 08/06/2018

Pour la Direction Générale Énergie Air et Climat du Ministère de la Transition écologique et solidaire	Sous-direction du système électrique et des énergies renouvelables, Bureau des énergies renouvelables.	Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE MINISTÈRE DE LA TRANSITION ÉCOLOGIQUE ET SOLIDAIRE

Correspondant commercial : **Météo-France, D2C, Toulouse**	Christophe.jacolin@meteo.fr + 33 5 61 07 86 85	METEO FRANCE
Correspondant technique : **Météo-France, DIR Nord, Villeneuve d'Ascq**	Julien.perfettini@meteo.fr + 33 3 20 67 66 31	

- page laissée intentionnellement vide -

Évolutions successives

Référence	Date	Version
Extension de la série d’observations horaires	23/04/2018	V1
Extension de la série d’observations horaires	08/06/2018	V2

Signatures

	Nom	Service
Rédacteur(s)	Valentine CHATEL	DIRN/EC
Relecteur(s)	Béatrice POUPONNEAU	DSM/EC/ECGC
Approbateur(s)	Julien PERFETTINI	DIRN/EC/D

Table des matières

1 Expression du besoin....8
1.1 Livrables....8
2 Problématique....8
2.1 Les données d’entrée....9
2.2 Les données de sortie....9
3 Méthodologie d'extension de la série temporelle d’observations horaires....9
3.1 Phase exploratoire des données d’entrée....9
3.1.1 Sélection du point AROME de référence....10
3.1.2 Identification des variables explicatives....11
3.2 Modélisation statistique de la force du vent (FF) à 100 m....11
3.2.1 Les étapes de l’apprentissage....14
3.2.2 Estimation de l’erreur d’estimation de FF....14
3.3 Modélisation statistique de la direction du vent (DD) à 100m....17
4 Principaux résultats sur l’extension de la série temporelle d’observations horaires....18
4.1 Phase exploratoire....18
4.2 Force du vent à 100 m....24
4.2.1 Étude des modèles statistiques pour l’estimation de FF 100 m....24
4.2.2 Synthèse des résultats sur l’échantillon d’apprentissage....31
4.2.3 Comparaison des modèles sur l’échantillon test....31
4.3 Direction du vent (DD) à 100 m....37
4.3.1 Étude des modèles statistiques pour l’estimation de DD 100 m....37
4.3.2 Comparaison des modèles sur l’échantillon de test....38
4.4 Synthèse des choix de modélisation de DD et FF....41
4.5 Extension de la série DD et FF à 100m....41
5 Limites d’utilisation des données reconstituées....42
6 Livraison de la série temporelle d’observations horaires....43

Liste des illustrations

Illustration 1: Roses des vents issues du modèle AROME.....18
Illustration 2: Rose des vents issue des observations sur site.....18
Illustration 3: Processus de sélection des échantillons d'apprentissages et de tests.....19
Illustration 4: Corrélation entre la variable à expliquer FFmFFARO et les variables explicatives. En haut : sous forme de bulles. En bas : chiffré (0 : absence de corrélation, 1 ou -1 : forte corrélation) 20
Illustration 5: Nuages de points par paires de variables au point AROME de référence. Plus les nuages sont alignés sur la bissectrice, plus il existe une corrélation linéaire entre les deux variables. De haut en bas : FFmFFARO, FFARO_10, FFARO_100, TARO_2, TARO_100, TARO_500, SQRT_TKEARO_10, SQRT_TKEARO_100, PMER ARO, HUARO_2, TPW850HPA_ARO.....21
Illustration 6: Histogramme du vent observé et des variables explicatives au point AROME. De haut en bas et de gauche à droite : FF, FFmFFARO, FFARO_10, FFARO_100, TARO_2, TARO_100, TARO_500, SQRT_TKE_10, SQRT_TKE_100, HUARO_2, PMERARO, TPW850HPA_ARO.....22
Illustration 7: Corrélation entre les variables à expliquer UmUARO et VmVARO et les variables explicatives. En haut : sous forme de bulles. En bas : chiffré (0 : absence de corrélation, 1 ou -1 : forte corrélation).....23
Illustration 8: Arbre de décision binaire calculé sur l'échantillon complet.....25
Illustration 9: Qualité des différents modèles linéaires, évaluée par validation croisée sur l'échantillon d'apprentissage. De haut en bas: RMSE, BIAIS, ECT, MAE . De gauche à droite : BRUT, LM0, LM, LM.STEP, GLM, GLM.STEP.....27
Illustration 10: Critères d'importances calculés sur l'échantillon complet (à gauche, le % de MSE, à droite le critère de pureté).....29
Illustration 11: Qualité des différentes forêts aléatoires, évaluée par validation croisée sur l'échantillon d'apprentissage. De haut en bas: RMSE, BIAIS, ECT, MAE . De gauche à droite : RF20 (A), RF50 (B), RF100 (C), RF200 (D) et RF500 (E).....30
Illustration 12 : De haut en bas : RMSE, BIAIS, ECT et MAE des erreurs de prévisions du modèle statistique. De gauche à droite, les modèle suivants : BRUT, GLM.STEP et RF100.....32
Illustration 13 : De haut en bas : les 3 seuils de dépassements de force de vent prévu / observés (3, 9 et 12 m/s) . A gauche le score PSS, à droite le score de fausse alarme (FA). Pour chaque figure, de gauche à droite, les modèle suivants : BRUT, GLM.STEP et RF100.....33
Illustration 14: QQ-plot (courbe de fiabilité), à gauche pour le modèle linéaire GLM.STEP, à droite pour le modèle de forêt aléatoire RF100.....34
Illustration 15 : Cycle annuel de la force du vent, sous forme de boîte à moustache. En noir : l'observation entre le 24/11/2016 et le 23/11/2017. En rouge : l'estimation de la force du vent issue du modèle GLM.STEP sur la même période. En bleu : l'estimation de la force du vent issue du modèle RF100 sur la même période. En vert : le modèle BRUT AROME sur la même période.....35
Illustration 16 : Cycle diurne de la force du vent, sous forme de boîte à moustache. En noir : l'observation entre le 24/11/2016 et le 23/11/2017. En rouge : l'estimation de la force du vent issue du modèle GLM.STEP sur la même période. En bleu : l'estimation de la force du vent issue du modèle RF100 sur la même période. En vert : le modèle BRUT AROME sur la même période.....36
Illustration 17: Gauche : Rose des vents observées pendant la campagne de mesure toutes forces de vents (> 3 m/s) confondues. Droite : Rose des vents issues du modèle AROME sur la même période.....37
Illustration 18: Roses des vents après reconstitution de la direction DD à l'aide des composantes U et V sur l'échantillon test, toutes forces de vents (> 3 m/s) confondues. De gauche à droite et de haut en bas : observation (FF et DD OBS), BRUT (FF issue du l'extension (§4.2) et DD AROME), GLM (FF issue de l'extension (§4,2) et DD reconstitué à partir de U.GLM et V.GLM) et GLMSTEP (FF issue de l'extension (§4,2) et DD reconstitué à partir de U.GLMSTEP et V.GLMSTEP).....38
Illustration 19: Scores de bonne prévision (H), de fausse alerte (FA) et Pierce Skill Score (PSS).

De haut en bas : H (H OBS/BRUT et H OBS/GLMSTEP), FA (FA OBS/BRUT et FA OBS/GLMSTEP) et PSS (PSS OBS/BRUT et PSS OBS/GLMSTEP).....40
Illustration 20: Roses des vents sur la campagne de mesure. Gauche : Modèle AROME brut (FF AROME et DD AROME). Centre : Observation (FF OBS et DD OBS). Droite : Choix pour l'extension (FF GLMSTEP et DD AROME).....41
Illustration 21: Rose des vents - 01/01/2000 00HTU au 23/11/2016 23HTU.....42

Liste des tableaux

Tableau 1: Score B95+ sur les 4 points modèle entourant le point d'observation.....18
Tableau 2: Description des 5 modèles statistiques linéaires.....26
Tableau 3: Scores par validation croisée sur l'échantillon d'apprentissage, et critère de BIC pour les différents modèles linéaires étudiés.....28
Tableau 4: Scores par validation croisée sur l'échantillon d'apprentissage pour les différentes forêt aléatoires étudiés.....31
Tableau 5: Scores par validation croisée sur l'échantillon test pour les différents modèles sélectionnés.....33
Tableau 6: Scores B95+ sur les roses des vents BRUT, GLM et GLMSTEP.....39

1 Expression du besoin

La Direction Générale de l'Énergie et du Climat (DGEC) a sollicité Météo-France pour la réalisation d'études de vent dans la zone d'implantation d'éoliennes en mer au large de Dunkerque. Le but étant d'identifier les risques susceptibles de se présenter dans les zones retenues comme propices au développement de l'éolien off-shore.

Ces études s'inscrivent dans le cadre du lancement du troisième appel d'offres éolien en mer conduit par la DGEC.

Ce rapport traite de l'extension de la série d'observations horaires issue de la campagne de mesure réalisée entre le 24/11/2016 0H TU et le 23/11/2017 23H TU sur le site situé au large de Dunkerque.

La période reconstituée est d'environ 17 années (01/01/2000 – 23/11/2016), calée sur la profondeur de la base climatologique AROME 2,5 km.

1.1 Livrables

Travaux	Livrable	Format
Extension temporelle des mesures de vent	Rapport d'étude complémentaire précisant la méthodologie et décrivant les résultats obtenus	Série de données

2 Problématique

La problématique principale de cette étude consiste à concevoir des modèles statistiques répondant au besoin ci-dessus exprimé par la DGEC.

Ce document s'attache à décrire la mise au point de ces modèles statistiques permettant l'extension de la série observée du lidar, installé au large de Dunkerque, entre le 24/11/2016 et le 23/11/2017 et à décrire les principaux résultats obtenus.

2.1 Les données d'entrée

Météo-France dispose d'un historique de 18 ans (2000-2017) de données à méso-échelle issues du modèle AROME à résolution 2,5 km et au pas de temps horaire. Le lecteur est invité à se reporter au rapport *Acquisition et suivi des mesures sur site durant un an*, livré en février 2018, pour plus de détails sur le modèle AROME.

2.2 Les données de sortie

Les données à livrer à la DGEC consistent en une série temporelle AAAAMMJJHH (année mois jour heure) horaire de vent (force et direction) à 100 m pour le point de mesure LIDAR au large de Dunkerque, sur la période du 01/01/2000 0H TU au 23/11/2016 23H TU.

3 Méthodologie d'extension de la série temporelle d'observations horaires

3.1 Phase exploratoire des données d'entrée

Tout travail de modélisation statistique est précédé d'une phase d'exploration des données. C'est une étape indispensable dont le but essentiel est d'aboutir à un sous-ensemble de données pertinentes pour la modélisation et d'analyser les liens entre ces variables.

Notamment concernant les travaux de ce rapport, il s'agit :

- d'identifier le point AROME le plus à même d'apporter en entrée des modèles statistiques un comportement le plus juste du vent sur le point d'observation,
- d'identifier les variables ou facteurs pouvant expliquer les variations de la force et de la direction du vent au cours du temps, pour les intégrer dans les modèles statistiques.

L'arbre binaire de décision peut être un support à la sélection de variables.

En sortie de cette analyse, nous disposons de la sélection des variables – dites **variables X^i explicatives** – qui vont être utilisées dans la recherche du modèle statistique ajustant le mieux les données observées.

3.1.1 Sélection du point AROME de référence

La **sélection du point AROME** de référence s'appuie sur quatre indicateurs de qualité dits B95+, présentés ci-dessous.

Soit f^i_o la rose des vents climatologique de référence et f^i_m la rose des vents du modèle AROME. Ces roses fréquentielles sont de 18 secteurs et 4 classes de vent ([0,1.5[,[1.5,4.5[,[4.5,8[,>8).

On considère

1. le critère principal C1 où la comparaison se fait sur l'ensemble des classes :

$$C1 = 100 - 0.5 * \sum_{i=1}^{18*4} |f_o^i - f_m^i|$$

2. le critère C2, pour lequel les classes de vitesse sont regroupées et on ne tient pas compte des vents calmes (<1.5m/s, pour lesquels on sait qu'il y a de grosses incertitudes instrumentales sur la direction) ce qui permet d'étudier la qualité de la modélisation des fréquences de direction :

$$C2 = 100 - 0.5 * \sum_{i=1}^{18} |\sum_{cl=1}^{3} f_o^i - \sum_{cl=1}^{3} f_m^i|$$

3. le critère C3 pour lequel les classes de direction sont regroupées ce qui permet d'étudier la qualité de la modélisation des fréquences de force du vent :

$$C3 = 100 - 0.5 * \sum_{i=1}^{4} |\sum_{secteur=1}^{18} f_o^i - \sum_{secteur=1}^{18} f_m^i|$$

Ces critères de qualité ont pour principal avantage leur simplicité d'interprétation, en particulier si le critère global est mauvais, on peut rapidement détecter si l'erreur vient d'un biais en direction ou d'un biais en force du vent. Par contre, il présente le défaut d'être très dépendant du nombre de classes: plus on a de classes, plus le critère est sévère. Dans notre cas, nous travaillons sur 18 classes en direction (tous les 20 degrés), et 4 classes en force du vent. Ceci rend le critère de direction très exigeant: un biais de 20° (possible lorsque la résolution horizontale du modèle n'est pas suffisante pour représenter correctement le relief) dans la direction du vent modèle, donnera un critère de direction nulle.

Enfin le dernier critère est la corrélation circulaire, appliqué uniquement sur la direction du vent. Le coefficient de corrélation est calculé comme la corrélation de Pearson pour deux variables linéaires X et Y. Dans la formule de calcul, cependant, (xi - xbar) et (yi - ybar) sont remplacés par sin (xi - xbar) et sin (yi - ybar), où xbar et ybar sont les directions moyennes des échantillons dans la seconde expression.

3.1.2 Identification des variables explicatives

Le choix des variables explicatives est guidé par la connaissance météorologique de l'origine du vent. Des variables complémentaires sont proposées pour intégrer certains comportements météorologiques, non directement accessible en sortie brute du modèle : prise en compte de certains cycles (diurne, saisonnier, …). Les variables suivantes ont été présélectionnées pour participer à la modélisation statistique :

- **HH** : prise en compte du cycle diurne (heure, sous forme de 3 facteurs correspondant à 3 plages horaires [4H – 10H], [11H - 17H] et [18H – 3H]),
- **MM** : prise en compte du cycle saisonnier (mois, variable catégorielle à 12 facteurs, de 1 à 12),
- Prise en compte d'éléments de stabilité de l'atmosphère :
  - **TPW850**: la température pseudo-adiabatique potentielle du thermomètre mouillé AROME à 850 hPa. Cette variable caractérise une masse d'air. C'est une variable quantitative,
  - **HU2M** : l'humidité AROME relative à 2 m – variable quantitative,
  - **SQRTKE** : la turbulence AROME à plusieurs niveaux – On exploite la racine carrée de cette variable, de manière à privilégier un lien linéaire avec la force du vent de l'observation – variable quantitative,
- Prise en compte de la situation météorologique générale :
  - **PMER** : la pression AROME réduite au niveau de la MER – variable quantitative,
  - **FF** : la force du vent AROME à plusieurs niveaux - variable quantitative**,**
  - **SECTEURARO :** la direction du vent AROME à 100m par secteur de 20° - variable catégorielle à 18 facteurs,
  - **T :** la température AROME à plusieurs niveaux – variable quantitative.

3.2 Modélisation statistique de la force du vent (FF) à 100 m

Le but de la modélisation statistique est d'établir un lien statistique robuste entre les variables explicatives X^i et les observations réelles de la force du vent à 100 m correspondant à la **variable Y à expliquer** :

$$Y=f(X^1,\ldots,Xi,\ldots,X^n)$$

Le type des variables statistiques diffère selon l'espace dans lequel elles prennent leurs valeurs.

Elles peuvent être qualitatives à valeurs dans un ensemble de cardinal fini ou quantitatives à valeurs réelles voire fonctionnelles. Certaines méthodes de modélisation s'adaptent à tout type de variables explicatives tandis que d'autres sont spécialisées.

Enfin, si la variable Y à expliquer est qualitative, on parle de discrimination, classement ou reconnaissance de forme tandis que si la variable Y est quantitative – notre cas – on parle, par habitude, d'un problème de régression. Dans ce cas encore, certaines méthodes sont spécifiques (régression linéaire, analyse discriminante), c'est-à-dire qu'elles sont soumises à des hypothèses de travail qu'il s'agit de vérifier, tandis que d'autres s'adaptent sans modification profonde remettant en cause leur principe (arbres de décision, forêt aléatoire …).

Compte tenu du problème à résoudre – estimation de la force du vent à 100m – et de l'expérience de Météo-France en matière de famille de modèles pouvant répondre au besoin, nous avons pré-sélectionné les types de modèles suivants :

Arbre binaire de décision

Cette méthode statistique est basée sur un découpage, par des hyperplans parallèles aux axes, de l'espace engendré par les variables explicatives. Nommés initialement partitionnement récursif ou segmentation, les développements importants de Breiman et col. (1984) les ont fait connaître sous l'acronyme de CART : Classification and Regression Tree ou encore de C4.5 (Quinlan, 1993) dans la communauté informatique. L'acronyme correspond à deux situations bien distinctes selon que la variable à expliquer, modéliser ou prévoir est qualitative (discrimination ou en anglais classification) ou quantitative (régression, notre cas).

Les solutions obtenues sont présentées sous une forme graphique simple à interpréter, même pour des néophytes, et constituent une aide efficace pour l'aide à la décision.

Le paramètre de réglage de ce modèle est la règle permettant de décider qu'un nœud est terminal : il devient ainsi une feuille. Ce point est le plus délicat. Il correspond à la recherche d'un modèle parcimonieux. Un arbre trop détaillé, associé à une sur-paramétrisation, est instable et donc probablement plus défaillant pour la prévision d'autres observations.

Un graphique représente la décroissance ou éboulis de la déviance (ou du taux de mal classés) en fonction du nombre croissant de feuilles dans l'arbre. Quand l'amélioration du critère est jugé trop petite ou négligeable, on élague l'arbre au nombre de feuilles obtenues.

Cette méthode ne requière pas d'hypothèse sur les distributions des variables.

Nous utilisons ce modèle essentiellement dans un but exploratoire des données.

Référence : *https://www.math.univ-toulouse.fr/~besse/**Wikistat**/pdf/st-m-app-cart.pdf*

Modèle linéaire général

Cette méthode cherche à exprimer l'espérance d'une variable réponse Y (qui est équivalente à sa moyenne, ou plus précisément sa moyenne attendue) en fonction d'une combinaison linéaire des variables explicatives X^i et d'un terme d'erreur (*i.e.*, de bruit) non contrôlé qui doit impérativement suivre une distribution normale et de même variance.

Référence : *www.math.univ-toulouse.fr/~besse/**Wikistat**/pdf/st-m-modlin-mlg.pdf*

Forêt aléatoire

La méthode d'approche est différente et plus coûteuse que celle des arbres binaires de décision. Elle consiste en l'utilisation du hasard pour améliorer les performances d'algorithmes de faibles qualités de classification. Nous rajoutons ainsi une part d'aléatoire au cours de la construction des différents arbres qui seront alors agrégés ensembles pour former une forêt.

La forêt aléatoire se construit en concevant un arbre sur un sous-échantillon tiré aléatoirement (ou échantillon « out-of-bag »). Ensuite, pour chacun des arbres à construire, un sous-ensemble de $q \leq P$ variables explicatives est sélectionné aléatoire et servira à leur élaboration respective.

L'objectif de cette approche est de rendre les arbres construits plus indépendants entre eux, ce qui offre de meilleures performances lors de l'agrégation en forêt. L'approche possède l'avantage d'être très fructueuse en grande dimension et d'être simple à mettre en œuvre. L'utilisation de la forêt aléatoire permet également de s'affranchir de toute phase d'élagage et de tout problème lié à la multicolinéarité des variables.

Comme pour tout modèle construit par agrégation, il n'y a pas d'interprétation directe. Néanmoins des informations pertinentes sont obtenues par le calcul et la représentation graphique d'indices proportionnels à l'importance de chaque variable dans le modèle agrégé et donc de sa participation à la régression ou à la discrimination.

Deux critères sont ainsi proposés pour évaluer l'importance de la j-ème variable.

- Le premier (Mean Decrease Accuracy) repose sur une permutation aléatoire des valeurs de cette variable. Plus la qualité, estimée par l'erreur out-of-bag, de la prévision est dégradée par la permutation des valeurs de cette variable, plus celle-ci est importante. Une fois le b-ème arbre construit, l'échantillon out-of-bag est prédit par cet arbre et l'erreur estimée enregistrée. Les valeurs de la j-ème variables sont aléatoirement permutées dans l'échantillon out-of-bag et l'erreur à nouveau calculée. La décroissance de la qualité de prévision est moyennée sur tous les arbres et utilisée pour évaluer l'importance de la variable j dans la forêt. Il s'agit donc d'une mesure globale mais indirecte de l'influence d'une variable sur la qualité des prévisions.
- Le deuxième (Mean Decrease Gini) est local, basé sur la décroissance d'entropie ou encore la décroissance de l'hétérogénéité définie à partir du critère de Gini. L'importance d'une variable est alors une somme pondérée des décroissances d'hétérogénéité induites lorsqu'elle est utilisée pour définir la division associée à un nœud.

Le paramètre le plus important à fixer est celui du nombre d'arbres à générer.

Référence : *https://www.math.univ-toulouse.fr/~besse/**Wikistat**/pdf/st-m-app-agreg.pdf*

3.2.1 Les étapes de l'apprentissage

Les traitements s'enchaînent de façon assez systématique selon le schéma suivant et quel que soit le domaine d'application :

1. Extraction des données
2. Exploration des données pour la détection de valeurs aberrantes ou seulement atypiques, d'incohérences, pour l'étude des distributions des structures de corrélation, recherche de typologies, pour des transformations des données...
3. Partition de l'échantillon en fonction de sa taille et des techniques qui seront utilisées pour estimer une erreur de prévision en vue des étapes de choix de modèle, puis de choix et certification de méthode.
4. Pour chacune des méthodes considérées : modèle linéaire général, arbre binaire de décision, forêt aléatoire…
    - estimer le modèle pour une valeur donnée d'un paramètre (ou plusieurs) de complexité : nombre de variables, de feuilles, d'arbres...
    - optimiser ce paramètre (ou ces paramètres) en fonction de la technique d'estimation de l'erreur retenue: validation croisée (notre cas), approximation par pénalisation de l'erreur d'ajustement (critères de BIC).
5. Comparaison des modèles optimaux obtenus (un par méthode) par estimation de l'erreur de prévision sur l'échantillon test.
6. Choix de la méthode retenue en fonction de ses capacités de prévision, de sa robustesse mais aussi, éventuellement, de l'interprétabilité du modèle obtenu.
7. Ré-estimation du modèle avec la méthode, le modèle et sa complexité optimisée à l'étape précédente sur l'ensemble des données.
8. Exploitation du modèle sur la base complète et de nouvelles données.

3.2.2 Estimation de l'erreur d'estimation de FF

Tous les auteurs s'accordent pour souligner l'importance qu'il y a à construire des modèles parcimonieux quelle que soit la méthode utilisée. Toutes les méthodes sont concernées : nombre de variables explicatives, de feuilles, nombre d'arbres… .

L'alternative est claire, plus un modèle est complexe et donc plus il intègre de paramètres et plus il est flexible donc capable de s'ajuster aux données engendrant ainsi une erreur faible d'ajustement. En revanche, un tel modèle peut s'avérer défaillant lorsqu'il s'agit de prévoir ou généraliser, c'est-à-dire de s'appliquer à des données qui n'ont pas participé à son estimation.

Il s'agit donc de trouver un compromis entre complexité du modèle et erreur d'estimation.

Le premier critère d'estimation de notre erreur de modélisation à considérer est le critère de qualité d'ajustement du modèle sur l'échantillon observé. Nous le complétons par un ensemble de critères complémentaires, définis comme suit :

- **RMSE**: l'erreur quadratique moyenne mesure la distance entre le modèle et la référence. Plus il est proche de 0, plus le modèle est proche de la référence.

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2)}$$

- **BIAIS** : il caractérise l'erreur systématique du modèle ; plus il est proche de 0, plus le modèle est proche en moyenne des observations. Un BIAIS positif, signifie que le modèle surestime le paramètre considéré.

$$\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)$$

- **ECT** : l'écart type donne la précision du modèle. Plus il est proche de 0, meilleur est le modèle.

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2 - (\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i))^2}$$

  $RMSE^2 = BIAIS^2 + ECT^2$

- **MAE** : L'erreur absolue moyenne

$$\frac{1}{N}\sum_{i=1}^{N}|P_i - O_i|$$

- **PSS** : Peirce Skill Score

  Soit un événement E (par exemple FF> 5 m/s)

  Soit un échantillon d'observations et un modèle offrant la possibilité de prévoir cet événement, il est possible de construire la table de confusion suivante :

		observations	
		E est vrai sur l'observation	E est faux sur l'observation
prévisions	E est vrai sur la prévision	$n_{11}(s)$	$n_{10}(s)$
	E est faux sur la prévision	$n_{01}(s)$	$n_{00}(s)$

Taux de bonnes prévisions: $H= n_{11}(s)/(n_{11}(s)+ n_{01}(s))$

Taux de fausses alertes: **FA**$= n_{10}(s)/(n_{10}(s)+ n_{00}(s))$

PSS=H-FA

PSS est compris entre -1 et 1. Si ce score est supérieur à 0, le taux de bonnes prévisions est supérieur à celui des fausses alertes. Plus il est proche de 1, meilleur est le modèle. Il est particulièrement adapté, pour évaluer la capacité du modèle à prévoir les valeurs extrêmes.

Ces critères lorsqu'ils sont appliqués à l'échantillon qui a été utilisé pour construire le modèle statistique (échantillon d'apprentissage) ne peuvent être qu'une estimation biaisée, car trop optimiste, de l'erreur de prévision puisqu'elle est liée aux données qui ont servi à l'ajustement du modèle.

Néanmoins, pour un même type de modèle (test de différents modèles linéaires par exemple, en vue de sélectionner le meilleur), les critères de qualité seront évalués sur l'échantillon d'apprentissage (étape 4 du paragraphe 3.2.1). De manière à avoir une estimation de la dispersion de ces indicateurs de qualité, nous appliquons une estimation par validation croisée.

La validation croisée est d'un principe simple, efficace et largement utilisée pour estimer une erreur moyennant un surplus de calcul. L'idée est d'itérer l'estimation de l'erreur sur plusieurs échantillons de validation puis d'en calculer la moyenne. Notamment ceci est indispensable pour réduire la variance et ainsi améliorer la précision lorsque la taille de l'échantillon initial est trop réduite pour en extraire des échantillons de validation et test de taille suffisante.

```
ALGORITHME DE LA VALIDATION CROISÉE :
1: Découper aléatoirement l'échantillon d'apprentissage en K parts (K-fold) de tailles approximativement égales selon une loi uniforme ;
2: for k=1 à K do
      3: mettre de côté l'une des parties,
      4: estimer le modèle sur les K- 1 parties restantes,
      5: calculer l'erreur sur chacune des observations qui n'ont pas participé à l'estimation
6: end for
7: moyenner toutes ces erreurs pour aboutir à l'estimation par validation croisée.
```

Nous exploitons également les nuages de points, aussi appelé diagramme de dispersion, qui sont une représentation graphique d'une série statistique à deux variables (vent observé et vent prévu par le modèle statistique). Ils permettent d'observer la relation entre ces deux variables. La distribution générant le diagramme de dispersion le plus proche de la diagonale peut alors être considérée comme le meilleur choix.

Enfin, nous exploitons un ultime critère, qui est la courbe de fiabilité (**QQ-Plot**): le QQ-plot est une technique graphique qui vise à comparer deux fonctions de répartition en traçant les quantiles de l'une en fonction de ceux de l'autre (le FF observé versus le FF estimé par le modèle statistique par exemple). Il peut être utilisé afin de déterminer si deux séries de données obéissent à la même loi de probabilité, pour comparer les fonctions de répartition entre elles ou pour vérifier qu'un ensemble de données empiriques suit une certaine loi de probabilité théorique. La distribution générant le q-q plot le plus proche d'une ligne droite de pente 1 peut alors être considérée comme le meilleur choix.

3.3 Modélisation statistique de la direction du vent (DD) à 100m

Selon la même méthodologie que pour la modélisation statistique de FF, nous vérifions si l'approche par modèles statistiques appliquée à l'estimation de la direction du vent permet d'apporter de l'information.

Pour cela, la variable Y à expliquer (DD 100m) étant une variable circulaire, nous allons réaliser une régression selon les deux composantes du vent, une zonale U et une méridienne V et estimer ensuite la direction du vent à partir des estimations de U et V.

Le passage de (DD, FF) à (U, V) s'effectue de la façon suivante :

```
U=sin(DD*pi/180)*FF
V=cos(DD*pi/180)*FF
```

4 Principaux résultats sur l'extension de la série temporelle d'observations horaires

4.1 Phase exploratoire

Choix du point AROME de référence

Le point du modèle AROME qui va servir de point de référence pour les variables explicatives des modèles statistiques est sélectionné à partir de la capacité de la rose des vents de ce point à être le plus proche possible de la rose des vents observée sur la campagne de mesure lidar qui s'est déroulée du 24 octobre 2016 au 23 novembre 2017.

Pour le site de mesure au large de Dunkerque, le point numéro 3 présente de très bon indicateurs B95+ (C1, C2 et C3). La corrélation circulaire en DD est plutôt favorable au point 3 ou au point 4 compte tenu du très faible écart.

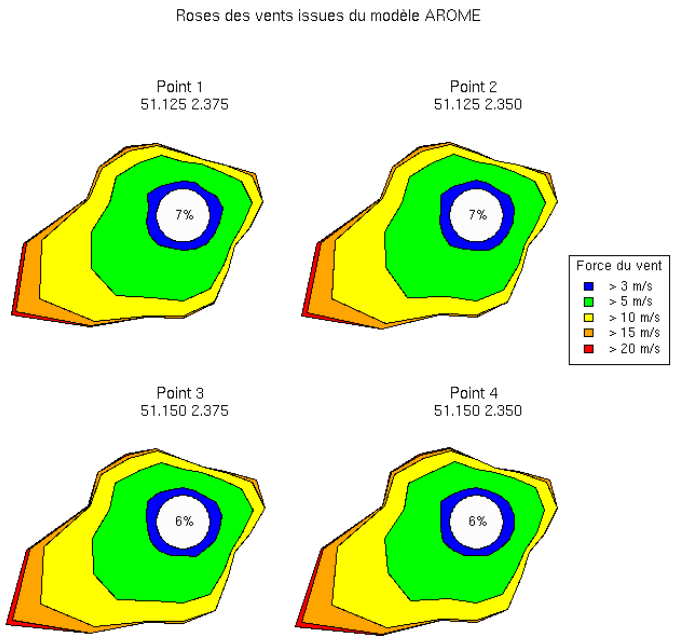


Illustration 1: Roses des vents issues du modèle AROME

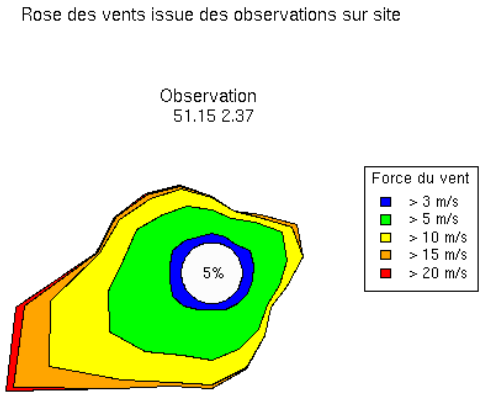


Illustration 2: Rose des vents issue des observations sur site

Point	Latitude	Longitude	DD&FF (C1)	DD (C2)	FF (C3)	Corrélation circulaire (C4)
1	51,125	2,375	87,06	93,51	89,83	93,81
2	51,125	2,350	87,30	94,14	90,11	94,05
3	**51,150**	**2,375**	**90,02**	**94,27**	**93,54**	**94,11**
4	51,150	2,350	90,59	94,28	94,56	94,24

Tableau 1: Score B95+ sur les 4 points modèle entourant le point d'observation.

Compte tenu de la proximité du point 3 du lidar et de la similarité entre la rose des vents sur site (illustration 2) et la rose des vents modèle (illustration 1), **on retient le point 3**.

Choix des échantillons d'apprentissage et de test

Pour la définition des modèles statistiques, nous devons séparer la série d'observation initiale en deux échantillons :

- l'échantillon d'apprentissage (on prend couramment 2/3 de la série initiale) pour ajuster les modèles statistiques,
- l'échantillon de test (on prend couramment le tiers restant) pour vérifier la robustesse du modèle.

L'échantillon initial étant petit, nous mettons en place une stratégie nous permettant de séparer de façon itérative la série d'observation initiale en deux échantillons (apprentissage et test). Les données seront alors tour à tour dans l'échantillon d'apprentissage et dans l'échantillon de test. Le processus est décrit dans l'illustration 3.

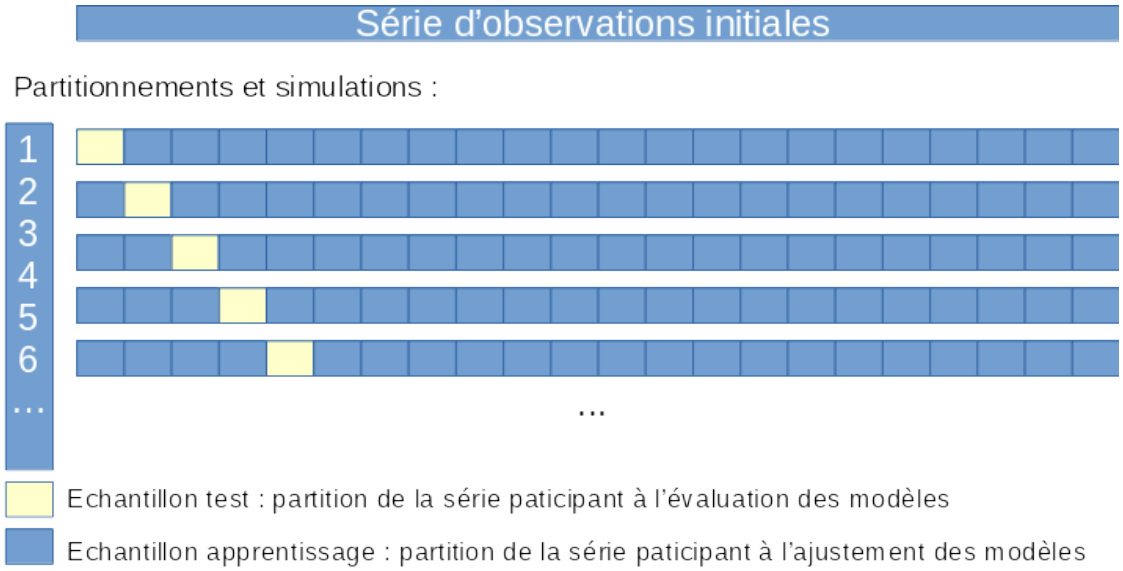


Illustration 3: Processus de sélection des échantillons d'apprentissages et de tests

Sélection des variables explicatives de la force du vent

De manière à mieux modéliser les valeurs extrêmes (retour d'expérience de nos premiers tests d'extension de série d'observations), nous avons cherché à **ajuster** des modèles statistiques, non plus sur la force FF du vent observé, mais sur **l'écart de cette force du vent avec la force du vent AROME** :

$Y=FF_{obs}-FF_{AROME}$ =**FFmFFARO**

L'ensemble des variables identifiées au paragraphe 3.1.2 est analysé.

Il s'agit d'examiner les liens entre la variable à expliquer Y et les variables explicatives $X^{i,}$ ou entre les variables explicatives, pour s'assurer qu'il y ait assez de corrélation entre chaque X^i et Y, et éviter (pour la modélisation linéaire) toute corrélation forte entre variables X^i (problème de colinéarité).

L'illustration 4 nous montre que la variable à expliquer est peu corrélée linéairement avec les variables explicatives. Cependant, certaines variables explicatives sont fortement corrélées entre-elles (exemple : TARO_2 et TARO_100). Au sein des nombreux niveaux disponibles pour les paramètres vent, température et turbulence, nous avons choisi de garder entre deux à trois niveaux par paramètres avec des coefficients de corrélations inférieurs à 0,98.

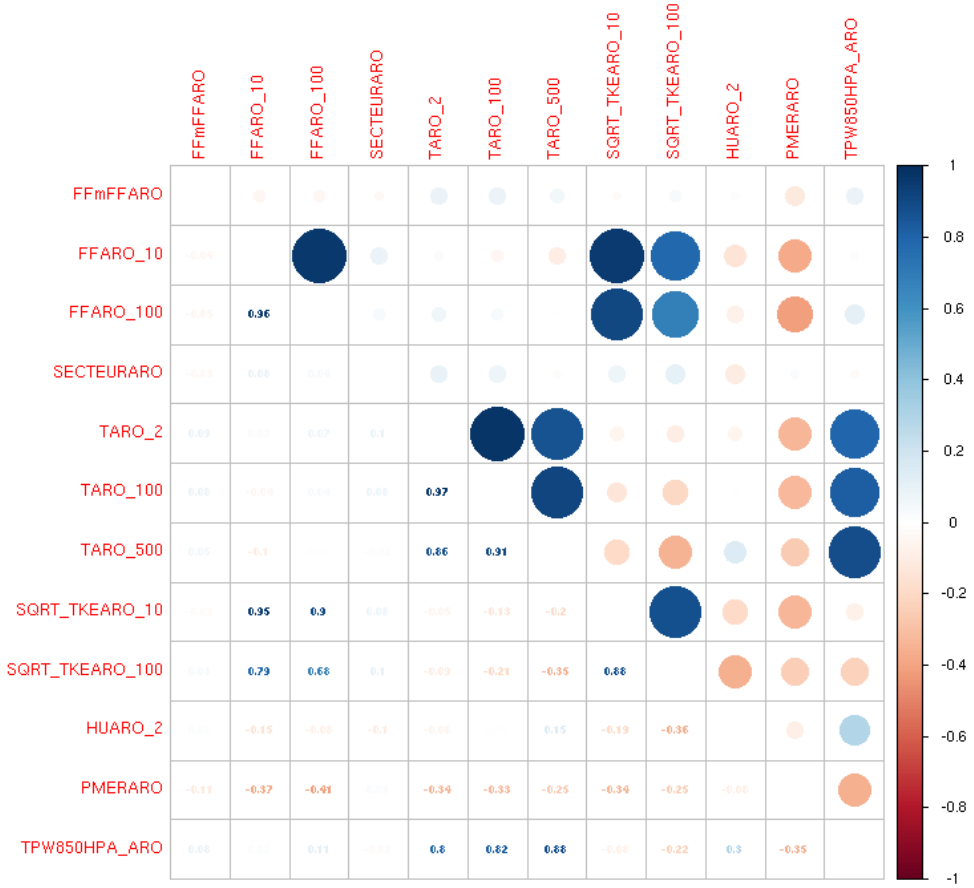


Illustration 4: Corrélation entre la variable à expliquer FFmFFARO et les variables explicatives. En haut : sous forme de bulles. En bas : chiffré (0 : absence de corrélation, 1 ou -1 : forte corrélation)

L'illustration 5 reprend les résultats précédents sous forme de nuages de points. Les variables explicatives sont fortement corrélées entre-elles mais leurs distributions ne sont pas exactement alignées sur la bissectrice, nous gardons donc toutes les variables présentées dans cette illustration.

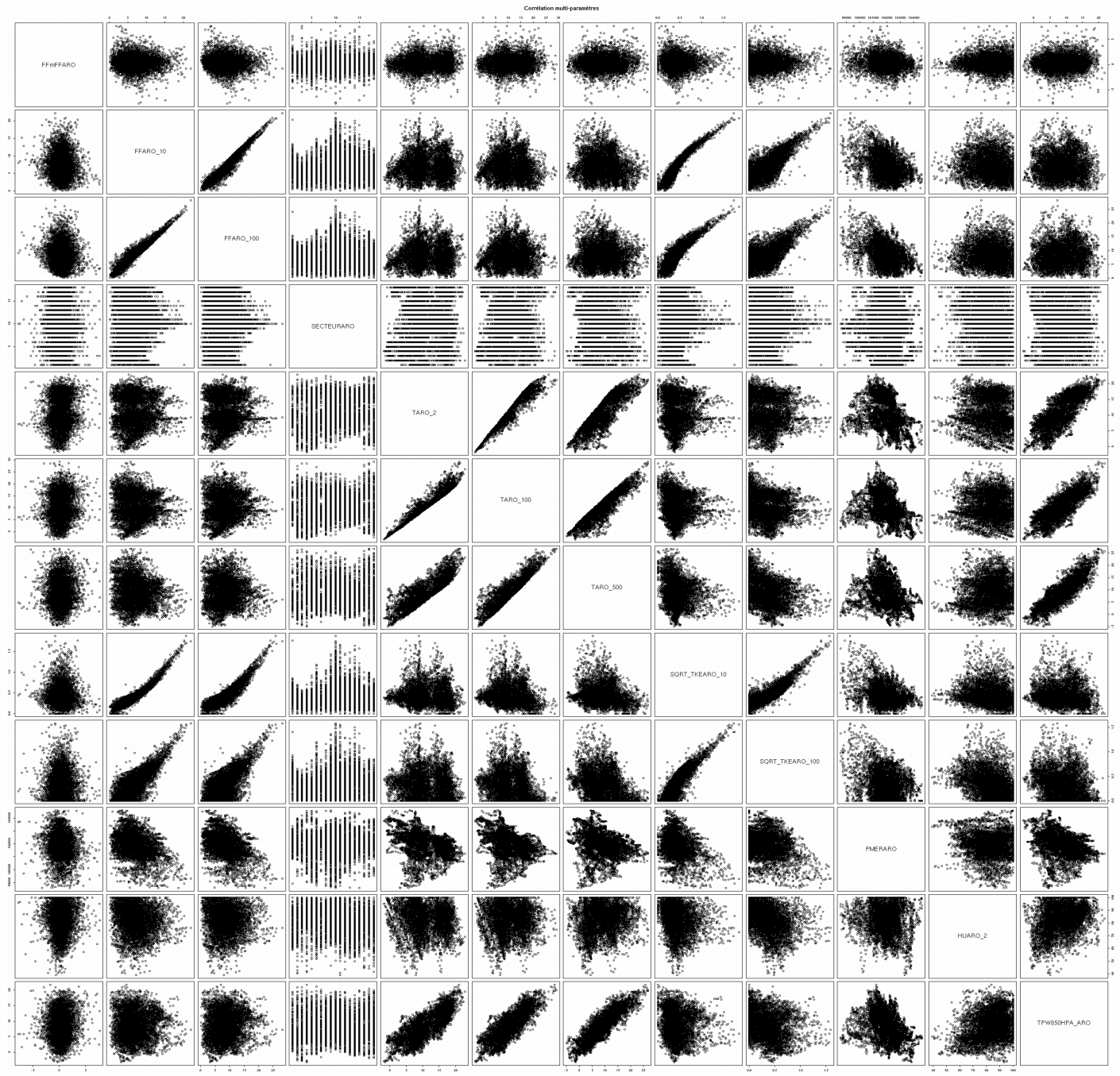


Illustration 5: *Nuages de points par paires de variables au point AROME de référence. Plus les nuages sont alignés sur la bissectrice, plus il existe une corrélation linéaire entre les deux variables. De haut en bas : FFmFFARO, FFARO_10, FFARO_100, TARO_2, TARO_100, TARO_500, SQRT_TKEARO_10, SQRT_TKEARO_100, PMER ARO, HUARO_2, TPW850HPA_ARO.*

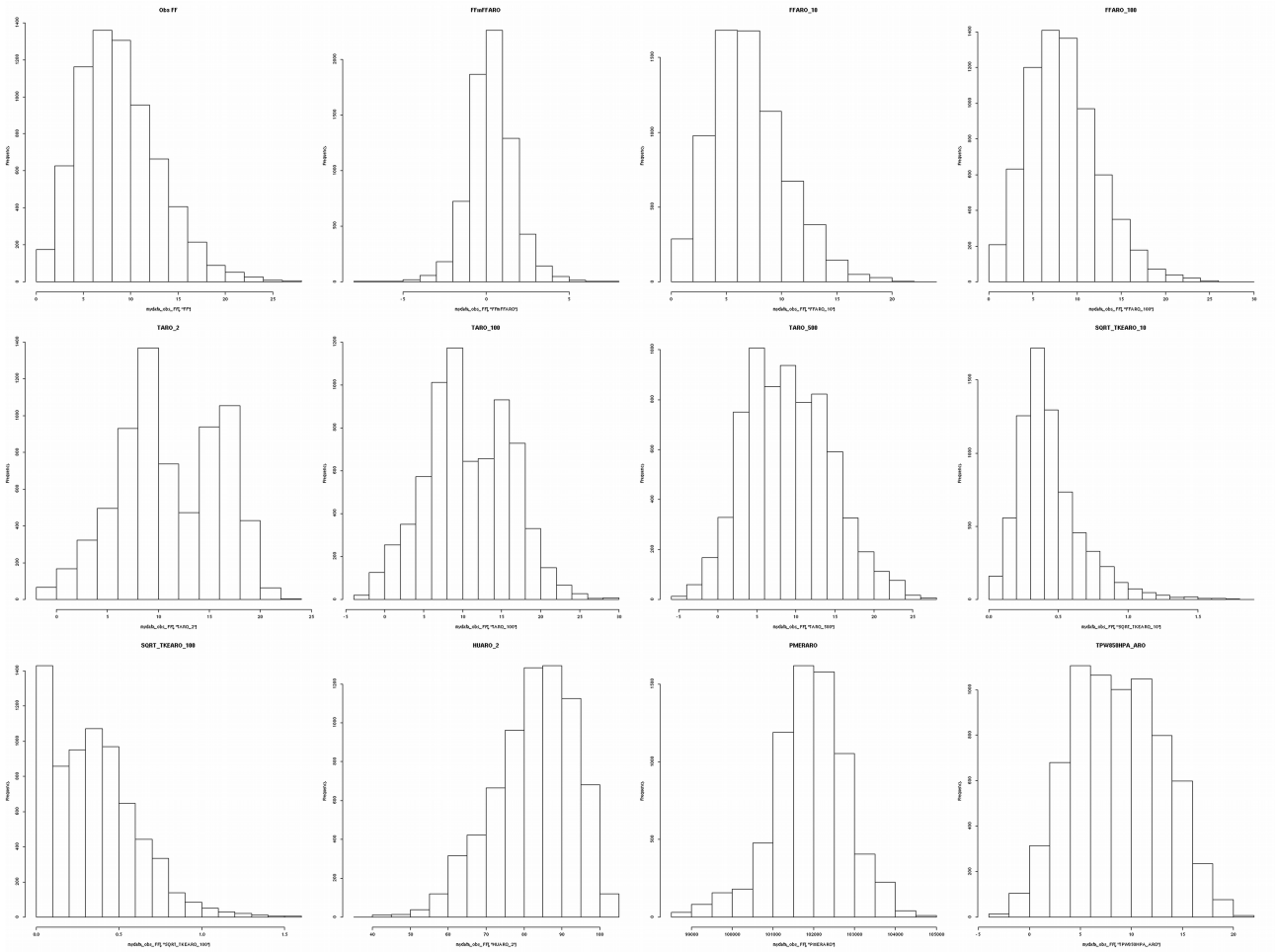


Illustration 6: Histogramme du vent observé et des variables explicatives au point AROME. De haut en bas et de gauche à droite : FF, FFmFFARO, FFARO_10, FFARO_100, TARO_2, TARO_100, TARO_500, SQRT_TKE_10, SQRT_TKE_100, HUARO_2, PMERARO, TPW850HPA_ARO

On observe qu'il n'y a pas de valeurs très extrêmes. Donc il y a peu de risque de « tirer » les régressions linéaires vers des valeurs peu représentatives.

Sélection des variables explicatives de la direction du vent

Pour les mêmes raisons que lors de la sélection des variables explicatives de la force du vent, nous avons cherché à **ajuster** des modèles statistiques, non sur la direction du vent observé, mais sur **l'écart de cette direction avec la direction du vent AROME** (nous ne pouvons pas étudier directement la direction du vent qui est une variable circulaire, nous devons donc étudier les composantes U et V (cf § 3.3) :

$Y=U_{obs} – U_{AROME}$ = **UmUARO** et $Y=V_{obs} – V_{AROME}$ = **VmVARO**

De façon similaire, il s'agit d'examiner les liens entre la variable à expliquer Y et les variables explicatives $X^{i,}$.

L'illustration 7 nous montre que les variables à expliquer UmUARO et VmVARO sont peu corrélées linéairement avec les variables explicatives. Cependant, U.ARO.10 (resp. V.ARO.10) et U.ARO.100 (resp. V.ARO.100) sont fortement corrélés.

Nous choisissons de garder le niveau 100 m qui correspond au niveau recherché pour la modélisation. **Nous écartons donc les variables U.ARO.10 et V.ARO.10.**

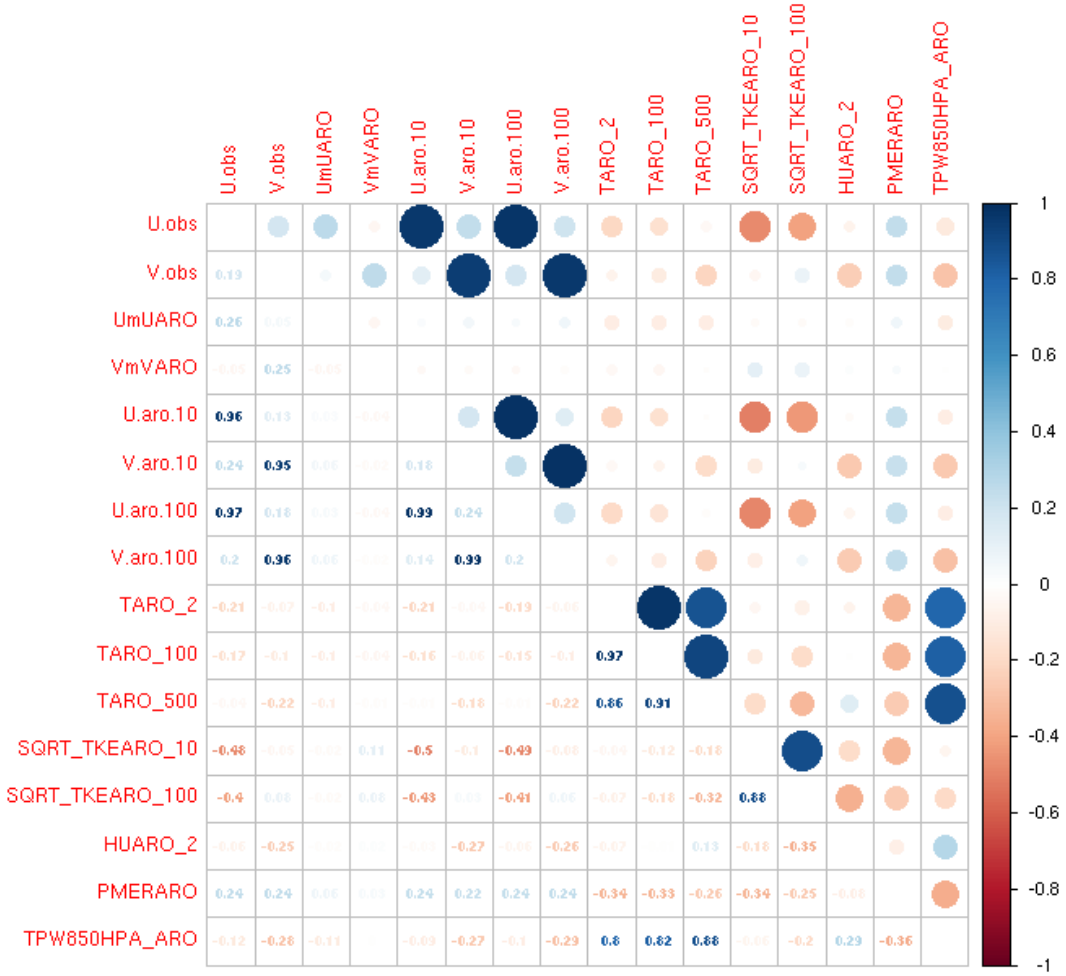


Illustration 7: Corrélation entre les variables à expliquer UmUARO et VmVARO et les variables explicatives. En haut : sous forme de bulles. En bas : chiffré (0 : absence de corrélation, 1 ou -1 : forte corrélation)

4.2 Force du vent à 100 m

4.2.1 Étude des modèles statistiques pour l'estimation de FF 100 m

Les arbres binaires de décision

Variable à expliquer : FFmFFARO

Variables explicatives : RegHH, MM, TPWARO850, PMERARO, HUARO2M, SQRTTKEARO (10 et 100), FFARO (10 et 100), TARO (2, 100 et 500) et SECTEURARO.

L'arbre binaire ne permet pas une prévision continue de la variable à expliquer Y. Il a surtout valeur de compréhension du rôle de chaque variable et de leur contribution en cascade dans la classification des éléments finaux de l'arbre.

On part d'un arbre avec le plus de feuilles possible, puis on va chercher à optimiser sa taille et à l'élaguer. On exploite toutes les variables explicatives à disposition.

Le facteur de complexité qui minimise l'erreur estimée sur l'échantillon d'apprentissage (par validation croisée) vaut **cp=0.002079242**. Une fois l'élagage réalisé, l'arbre complet dispose de 5 nœuds. C'est le modèle **TREE.OPTIM**.

On observe dès le haut de l'arbre (illustration 8), une séparation des données par la direction du vent. Les variables pression réduite au niveau de la mer, force du vent à 10m et force du vent à 100m ont également une grande importance et sont génératrices de séparations.

Importance des variables (ordre croissant d'importance de gauche à droite) :

REGHH	TPW ARO_ 850	HUARO _2	TARO_2	TARO_ 100	SQRT_ TKE ARO_ 100	MM	TARO_ 500	SQRT_ TKE ARO_10	FFARO _10	FFARO _100	PMER ARO	SECTE RU ARO
49,34	167,9	177,75	182,62	204,26	246,09	249,24	269,29	343,09	424,02	474,36	523,85	686,28

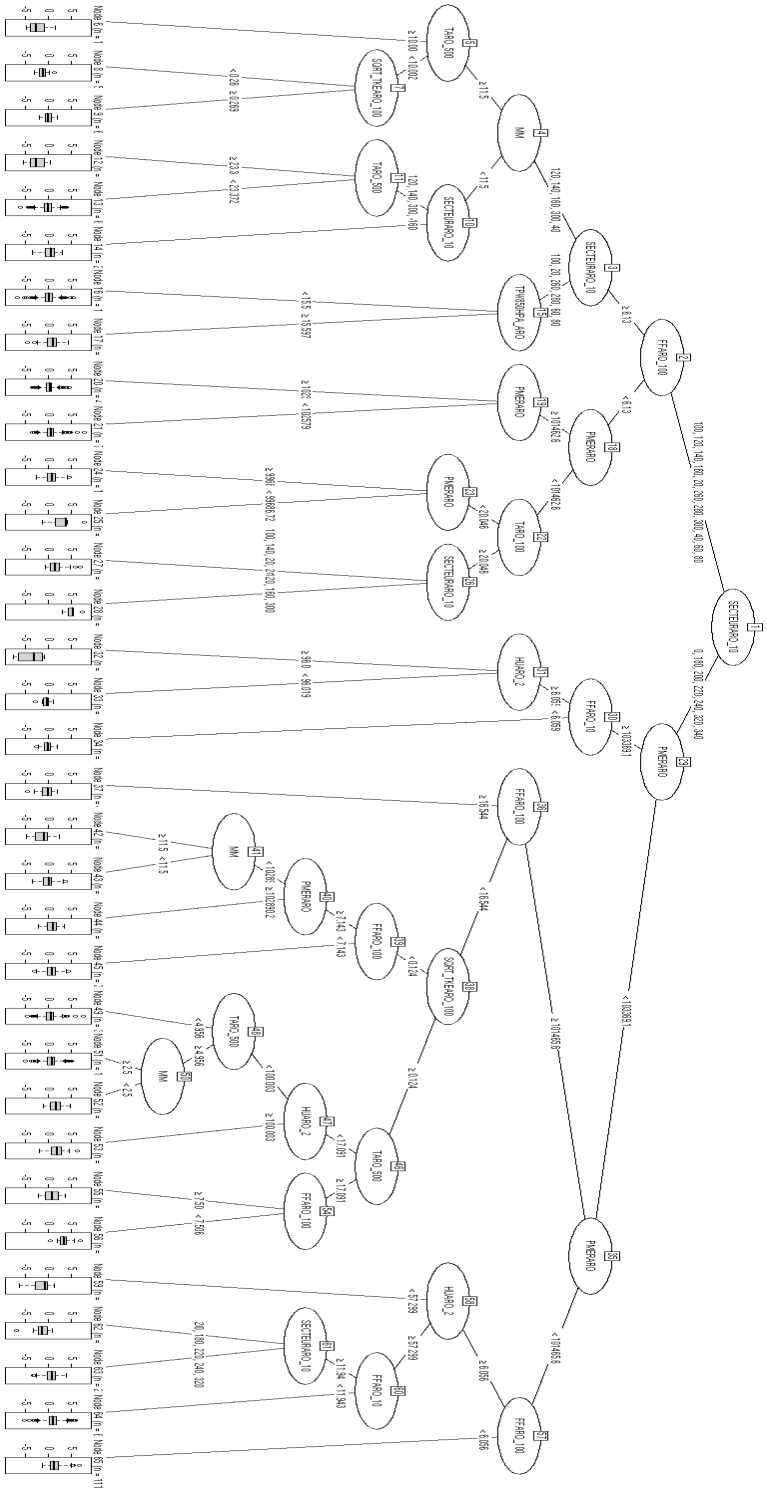


Illustration 8: Arbre de décision binaire calculé sur l'échantillon complet

Le modèle brut et les modèles linéaires

Variable à expliquer : FF (modèle brut) et FFmFFARO (modèles linéaires)

Variables explicatives : RegHH, MM, TPWARO850, PMERARO, HUARO2M, SQRTTKEARO (10 et 100), FFARO (10 et 100), TARO (2, 100 et 500) et SECTEURARO.

Nous avons étudié le modèle brut (FF=FFARO_100) et 5 modèles linéaires différents décrits ci-dessous :

Nom du modèle linéaire	Forme du modèle linéaire
LM0	`FFmFFARO=f(FFARO)`
LM	`FFmFFARO=f(REGHH, MM, FFARO_10, FFARO_100, TARO_2, TARO_100, TARO_500, SQRTTKEARO_10, SQRTTKEARO_100, TPWARO850, HUARO2, PMERARO, SECTEURARO)`
LM.STEP	(sélection automatique des variables significatives à partir du modèle LM ci-dessus)
GLM	`FFmFFARO=f(REGHH, MM, FFARO_10, FFARO_100, TARO_2, TARO_100, TARO_500, SQRTTKEARO_10, SQRTTKEARO_100, TPWARO850, HUARO2, PMERARO, SECTEURARO)^2` (interaction deux à deux)
GLM.STEP	(sélection automatique des interactions significatives à partir du modèle GLM ci-dessus)

Tableau 2: Description des 5 modèles statistiques linéaires

Les scores de qualité, obtenus sur l'échantillon d'apprentissage, de l'ajustement des modèles à l'observation sont présentées dans l'illustration 9.

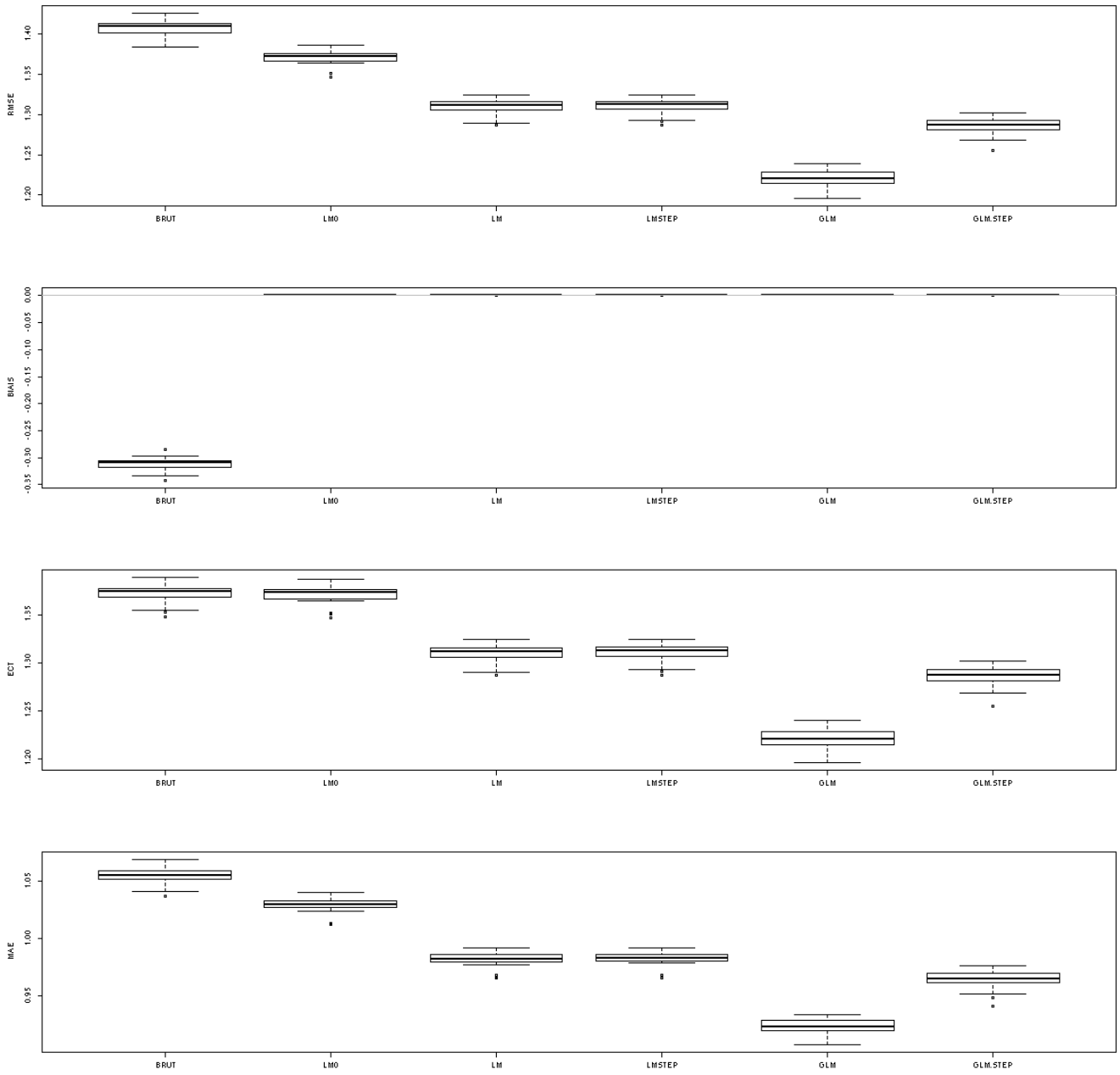


Illustration 9: Qualité des différents modèles linéaires, évaluée par validation croisée sur l'échantillon d'apprentissage. De haut en bas: RMSE, BIAIS, ECT, MAE . De gauche à droite : BRUT, LM0, LM, LM.STEP, GLM, GLM.STEP.

Modèles	RMSE	BIAIS	ECT	MAE	PSS1/FA1	PSS2/FA2	PSS3/FA3	BIC
BRUT	1,407	-0,312	1,372	1,055	0,639 / 0,345	0,788 / 0,106	0,865 / 0,053	/
LM0	1,371	-4,49 E-16	1,371	1,028	0,722 / 0,255	0,782 / 0,082	0,847 / 0,044	23252
LM	1,309	-5,94 E-15	1,309	0,981	0,75 / 0,226	0,792 / 0,082	0,859 / 0,042	22893
LM.STEP	1,31	-6,84 E-15	1,31	0,982	0,755 / 0,221	0,793 / 0,083	0,86 / 0,043	22861
GLM (avec interaction)	1,22	1,29 E-13	1,22	0,923	0,757 / 0,218	0,81 / 0,078	0,856 / 0,04	24524
GLM.STEP (avec interaction)	**1,285**	**9,72 E-14**	**1,285**	**0,964**	**0,757 / 0,217**	**0,798 / 0,081**	**0,86 / 0,04**	**22811**

Tableau 3: Scores par validation croisée sur l'échantillon d'apprentissage, et critère de BIC pour les différents modèles linéaires étudiés

Les résultats précédents montrent que la GLM donne les meilleurs résultats. Nous lui préférons la **GLM.STEP** (modèle linéaire avec sélection automatique des variables explicatives avec interaction) qui est proche en termes de résultat et moins complexe.

Les forêts aléatoires

Variable à expliquer : FFmFFARO

Variables explicatives : RegHH, MM, TPWARO850, PMERARO, HUARO2M, SQRTTKEARO (10 et 100), FFARO (10 et 100), TARO (2, 100 et 500) et SECTEURARO.

Nous agissons sur le nombre d'arbres dans la forêt aléatoire pour trouver l'arbre optimal en termes de qualité d'ajustement sur l'échantillon d'apprentissage, tout en restant un modèle parcimonieux.

L'importance des variables est la suivante:

RegHH	SQRT_TKEARO_10	SQRT_TKEARO_100	TARO_100	FFARO_10	MM	HUARO_2	FFARO_100	PMERARO	TPWARO_850	TARO_2	TARO_500	SECTEURARO
12,10	16,88	17,40	18,45	21,10	21,26	21,29	22,07	22,82	23,20	25,68	25,85	44,94

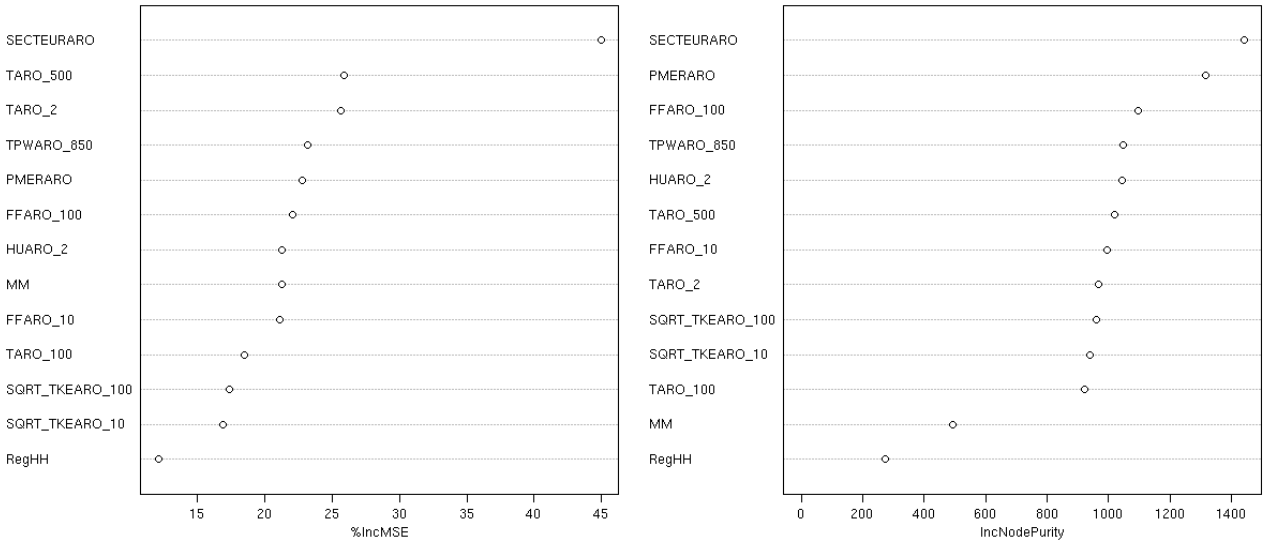


Illustration 10: Critères d'importances calculés sur l'échantillon complet (à gauche, le % de MSE, à droite le critère de pureté).

Conformément aux autres modèles (linéaires, arbres …), SECTEURARO joue un rôle prépondérant, en effet, il contribue nettement à diminuer la variance des résidus et contribue de façon importante au critère de pureté.

TARO_500 et TARO_2 jouent un rôle important dans ce modèle au niveau du % de MSE (MSE = RMSE au carré) mais sont plus en retrait sur le critère de pureté, tandis que PMERARO contribue de façon plus marquée au critère de pureté.

Inversement, REGHH n'explique que faiblement le % de MSE et intervient peu dans le critère de pureté.

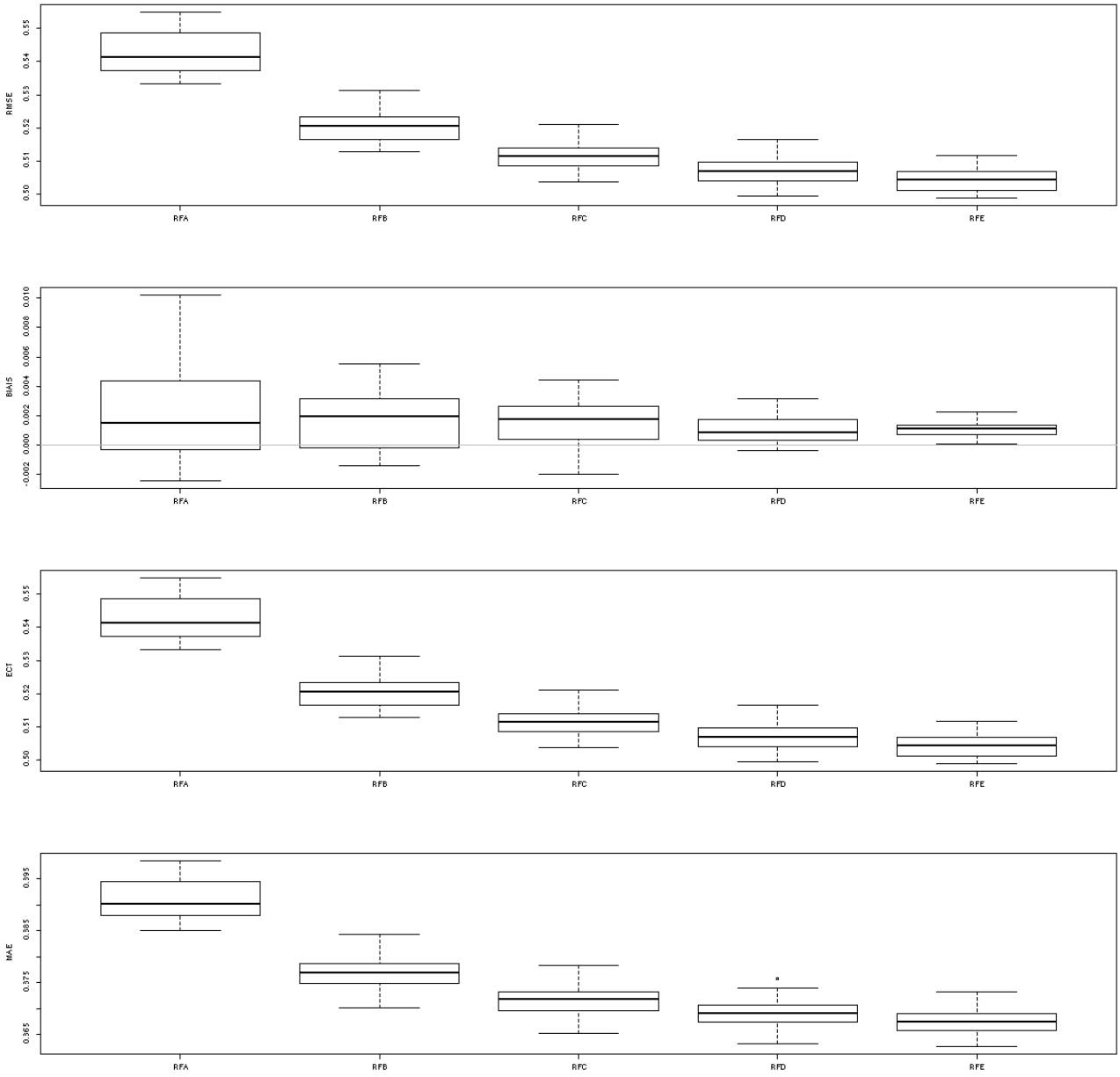


Illustration 11: Qualité des différentes forêts aléatoires, évaluée par validation croisée sur l'échantillon d'apprentissage. De haut en bas: RMSE, BIAIS, ECT, MAE . De gauche à droite : RF20 (A), RF50 (B), RF100 (C), RF200 (D) et RF500 (E).

Modèles	RMSE	BIAIS	ECT	MAE	PSS1/FA1	PSS2/FA2	PSS3/FA3
RF20	0,543	0,002	0,543	0,391	0,941 / 0,046	0,921 / 0,032	0,944 / 0,017
RF50	0,52	0,002	0,52	0,377	0,948 / 0,039	0,924 / 0,03	0,946 / 0,017
RF100	**0,512**	**0,0052**	**0,512**	**0,372**	**0,953 / 0,033**	**0,925 / 0,03**	**0,947 / 0,017**
RF200	0,507	0,001	0,507	0,369	0,955 / 0,032	0,926 / 0,03	0,948 / 0,016
RF500	0,504	0,001	0,504	0,368	0,959 / 0,028	0,926 / 0,03	0,948 / 0,016

Tableau 4: Scores par validation croisée sur l'échantillon d'apprentissage pour les différentes forêt aléatoires étudiés

Le nombre d'arbres conservés est 100 (nous observons très peu de différences sur les indicateurs de qualité entre 100, 200 et 500 arbres) : c'est le **modèle RF100.**

4.2.2 Synthèse des résultats sur l'échantillon d'apprentissage

A l'issue de cette première phase, nous avons sélectionné un modèle pour chaque famille de modèle statistique.

Il nous reste donc trois modèles : BRUT (modèle AROME), GLMSTEP (famille des modèles linéaires) et RF100 (famille des forêts aléatoires).

Le modèle TREE.OPTIM (arbre binaire de décision) ne sera pas utilisé pour la prévision, car il restitue une prévision par classe de vent.

4.2.3 Comparaison des modèles sur l'échantillon test

Dans ce paragraphe, nous conduisons une comparaison entre chaque modèle statistique sélectionné précédemment. L'objectif est de sélectionner le modèle final, qui sera utilisé pour l'extension de la série d'observations horaires.

On examine les résultats de la validation croisée sur l'échantillon test.

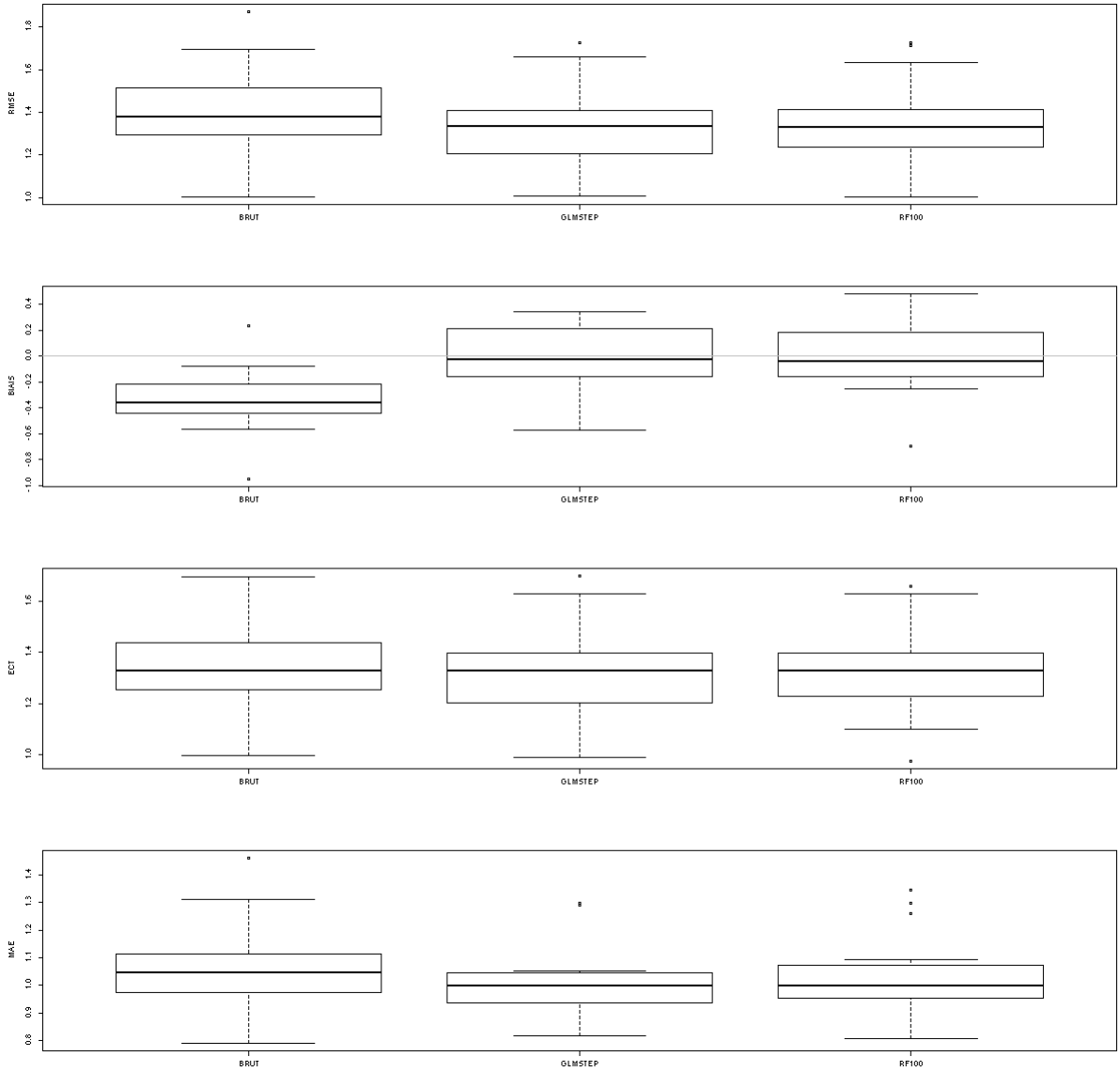


Illustration 12 : De haut en bas : RMSE, BIAIS, ECT et MAE des erreurs de prévisions du modèle statistique. De gauche à droite, les modèle suivants : BRUT, GLM.STEP et RF100.

Les indicateurs sont optimaux et très proches pour le modèle de forêt aléatoire à 100 arbres (RF100) et pour le modèle linéaire à interactions sélectionnées (GLM.STEP).

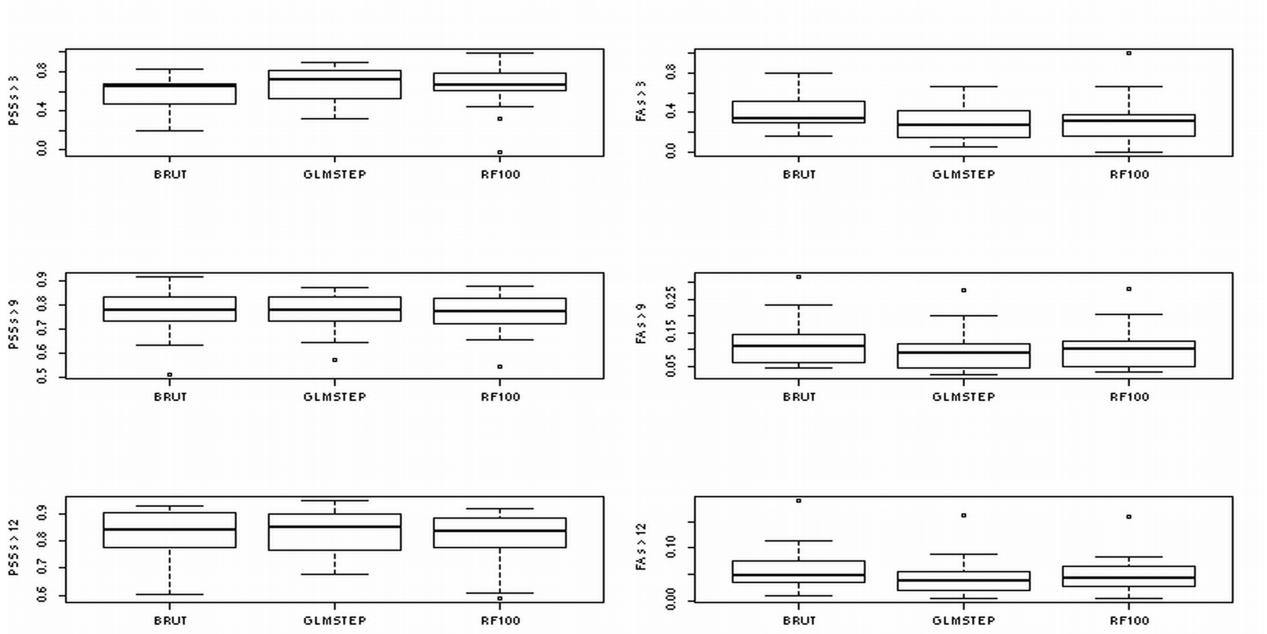


Illustration 13 : De haut en bas : les 3 seuils de dépassements de force de vent prévu / observés (3, 9 et 12 m/s) . A gauche le score PSS, à droite le score de fausse alarme (FA). Pour chaque figure, de gauche à droite, les modèle suivants : BRUT, GLM.STEP et RF100.

On cherche le PSS le plus élevé, et le FA le plus faible : globalement les modèles sont très proches. Le modèle de forêt aléatoire à 100 arbres (RF100) donne très légèrement de meilleurs résultats notamment pour le PSS > 3 m/s et le FA > 3 m/s.

Dans tous les cas, on montre que l'utilisation d'un modèle statistique améliore les sorties brutes d'AROME, si on compare les modèles au modèle le plus à gauche (BRUT).

Sur l'échantillon de test et par validation croisée, on obtient :

Modèle	RMSE	BIAIS	ECT	MAE	PSS1/FA1	PSS2/FA2	PSS3/FA3
BRUT	1,4	-0,331	1,343	1,061	0,574 / 0,409	0,767 / 0,114	0,830 / 0,059
GLMSTEP	**1,338**	**-0,003**	**1,317**	1,015	**0,649 / 0,320**	**0,770 / 0,093**	**0,830 / 0,046**
RF100	**1,347**	**-0,010**	**1,324**	**1,028**	**0,650 / 0,318**	**0,764 / 0,101**	**0,807 / 0,049**

Tableau 5: Scores par validation croisée sur l'échantillon test pour les différents modèles sélectionnés

Comme vu précédemment, les modèles RF100 et GLMSTEP sont proches, pour les différencier nous allons tester la robustesse de ces modèles sur l'échantillon test complet.

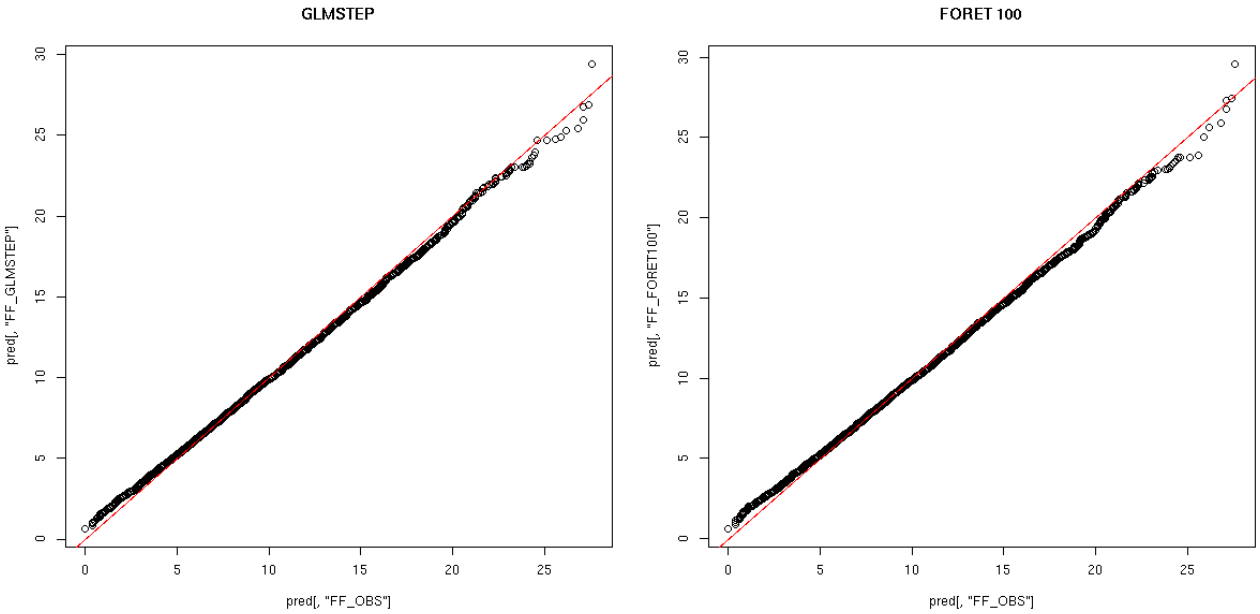


Illustration 14: QQ-plot (courbe de fiabilité), à gauche pour le modèle linéaire GLM.STEP, à droite pour le modèle de forêt aléatoire RF100.

La courbe de fiabilité est légèrement meilleure pour le modèle GLM.STEP que pour le modèle RF100.

La corrélation linéaire entre les observations et les estimations issues de GLM.STEP sur l'échantillon de test vaut 0,947 et la corrélation linéaire entre les observations et les estimations issues de RF100 sur l'échantillon de test vaut 0.946 (contre 0.945 entre l'observation et AROME)

Nous examinons la force du vent restituée par les deux modèles statistiques et par le modèle brut, en termes de cycle diurne et de cycle annuel et la comparons aux observations issues de la campagne de mesure lidar.

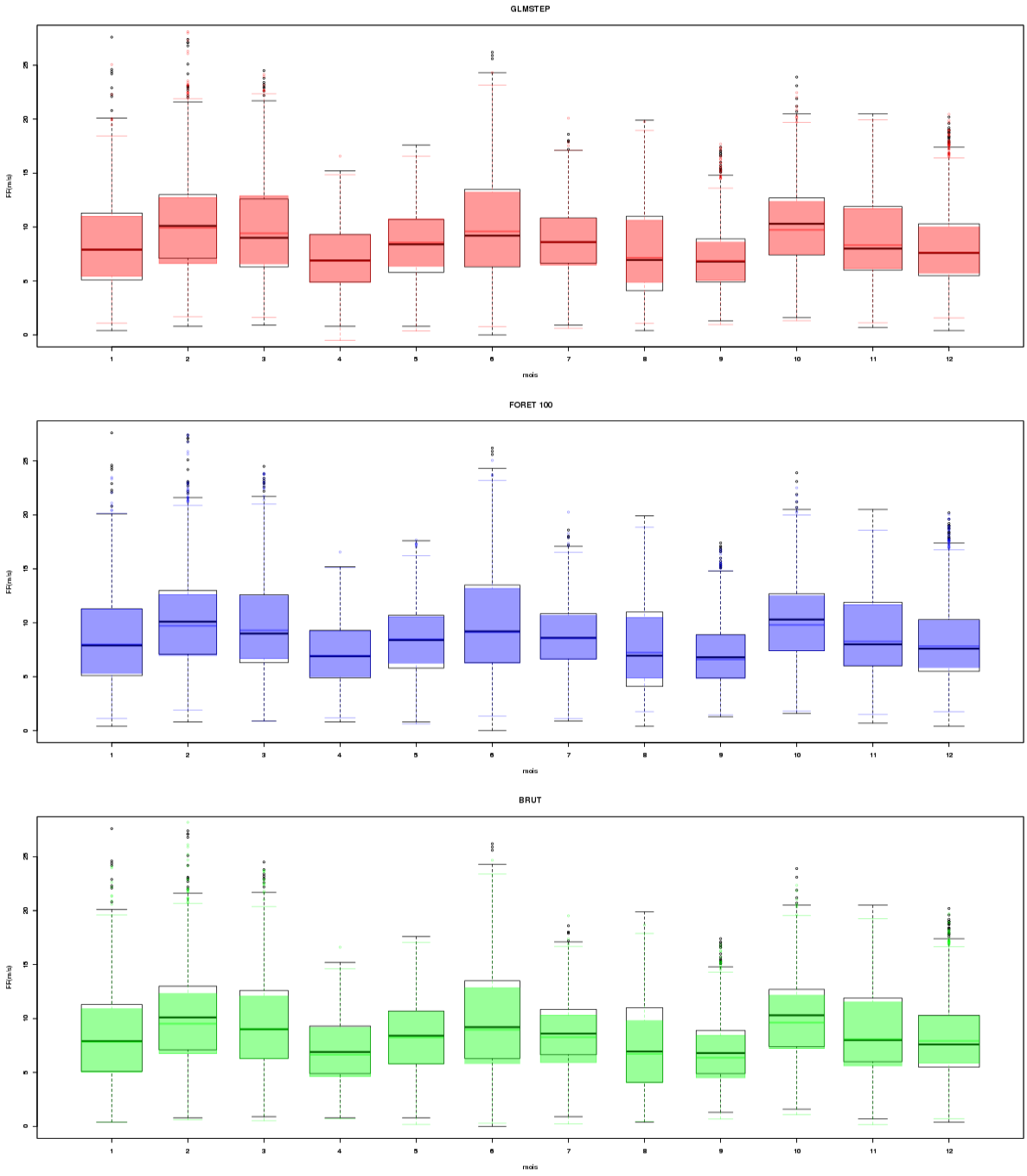


Illustration 15 : Cycle annuel de la force du vent, sous forme de boîte à moustache. En noir : l'observation entre le 24/11/2016 et le 23/11/2017. En rouge : l'estimation de la force du vent issue du modèle GLM.STEP sur la même période. En bleu : l'estimation de la force du vent issue du modèle RF100 sur la même période. En vert : le modèle BRUT AROME sur la même période.

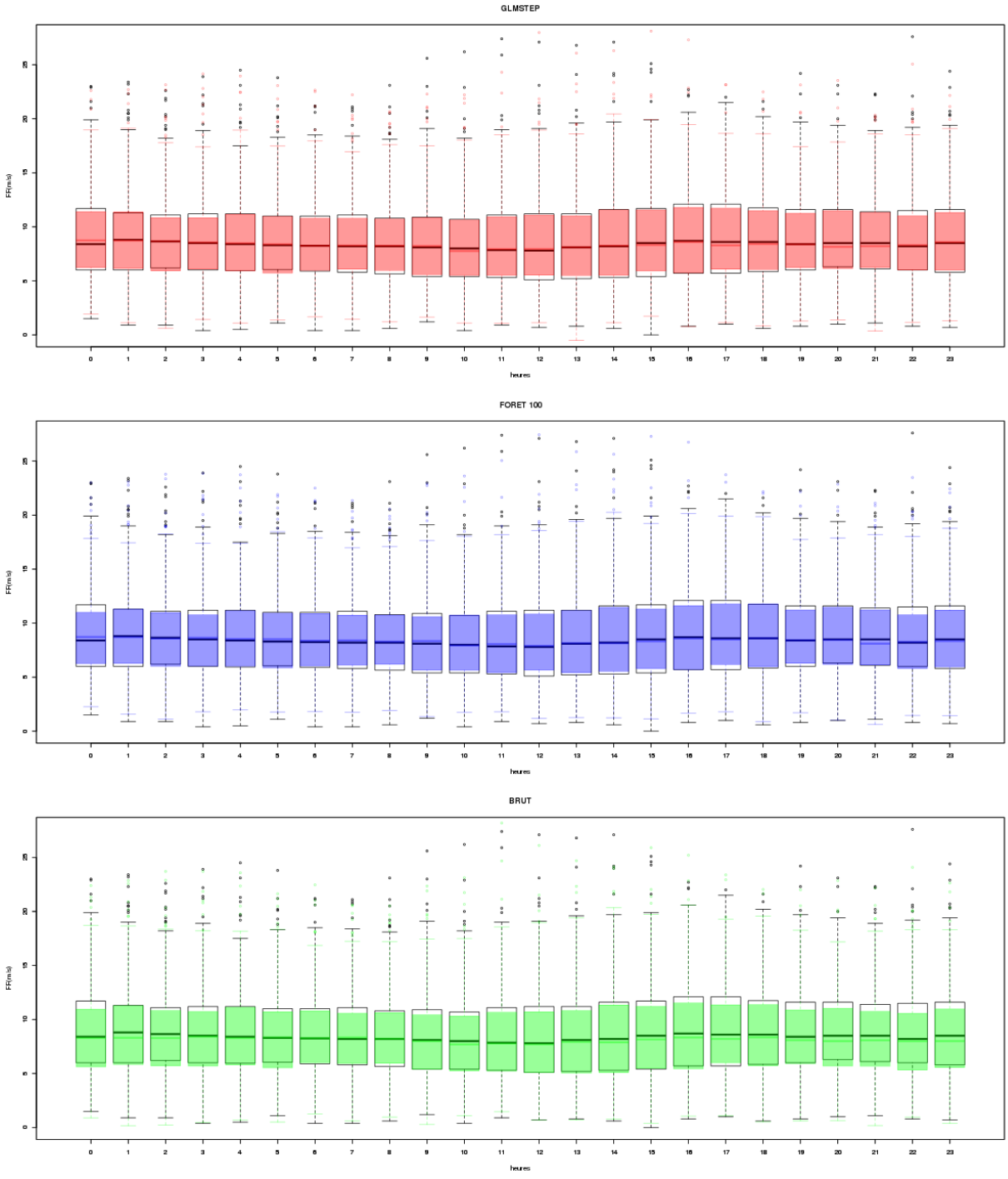


Illustration 16 : Cycle diurne de la force du vent, sous forme de boîte à moustache. En noir : l'observation entre le 24/11/2016 et le 23/11/2017. En rouge : l'estimation de la force du vent issue du modèle GLM.STEP sur la même période. En bleu : l'estimation de la force du vent issue du modèle RF100 sur la même période. En vert : le modèle BRUT AROME sur la même période.

Les modèles statistiques permettent de corriger les légers défauts de distribution des forces de vent au cours de l'année.

Nous privilégions la sélection du modèle GLM.STEP qui possède une meilleure courbe de fiabilité et qui restitue très bien les cycles de force du vent.

4.3 Direction du vent (DD) à 100 m

4.3.1 Étude des modèles statistiques pour l'estimation de DD 100 m

La phase exploratoire (§4.1) a montré une **très bonne correspondance modèle – observations** (cf score B95+ du tableau 1). Les roses des vents de l'illustration 17 présentent la comparaison entre l'observation et le point modèle issu du modèle AROME, toutes forces de vents supérieures à 3 m/s confondues, sur la campagne de mesure.

Les secteurs 20°, 160 et 220 – 260 sont les moins bien restitués par le modèle. Ce sont ces secteurs qui vont nous permettre de choisir le modèle le plus représentatif.

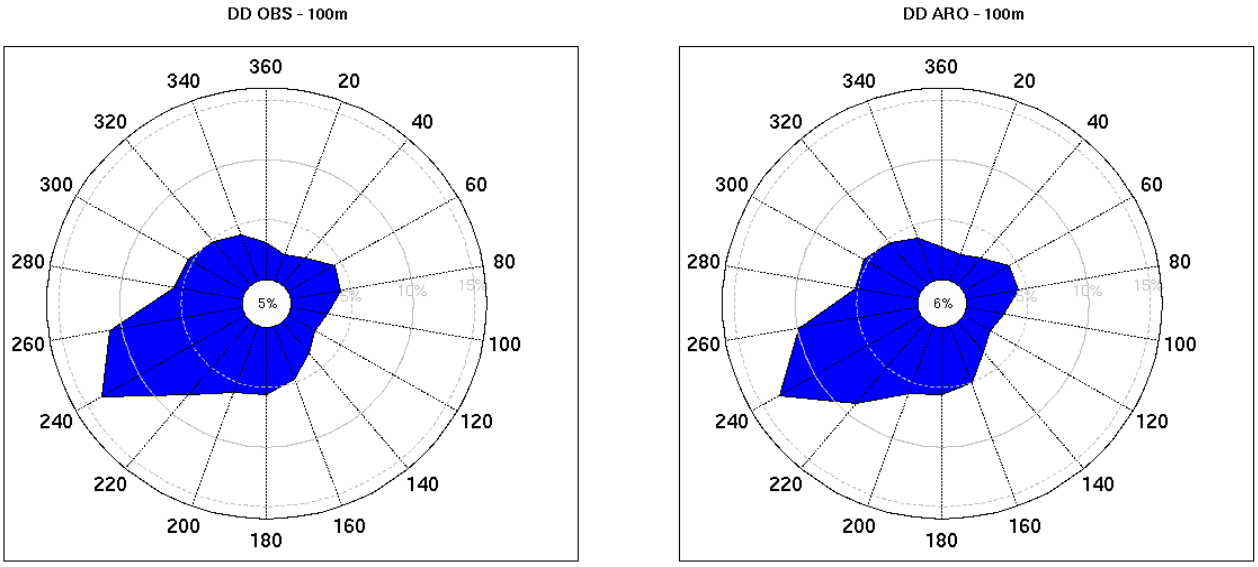


Illustration 17: Gauche : Rose des vents observées pendant la campagne de mesure toutes forces de vents (> 3 m/s) confondues. Droite : Rose des vents issues du modèle AROME sur la même période.

En appliquant, sur l'échantillon d'apprentissage, une démarche statistique similaire à celle utilisée pour la force du vent, pour chacune des composantes U et V du vent, il est très difficile de dégager un modèle par rapport aux autres sur les critères d'évaluations identifiés pour l'estimation statistique de FF.

Nous avons fait le choix de **sélectionner le modèle** sur sa **capacité à reconstituer correctement une rose des vents.**

4.3.2 Comparaison des modèles sur l'échantillon de test

Nous examinons à présent les roses des vents après reconstitution de la direction du vent à partir des composantes U et V (illustration 18) sur l'échantillon de test complet.La force du vent étant issue du modèle statistique GLMSTEP (§4.2).

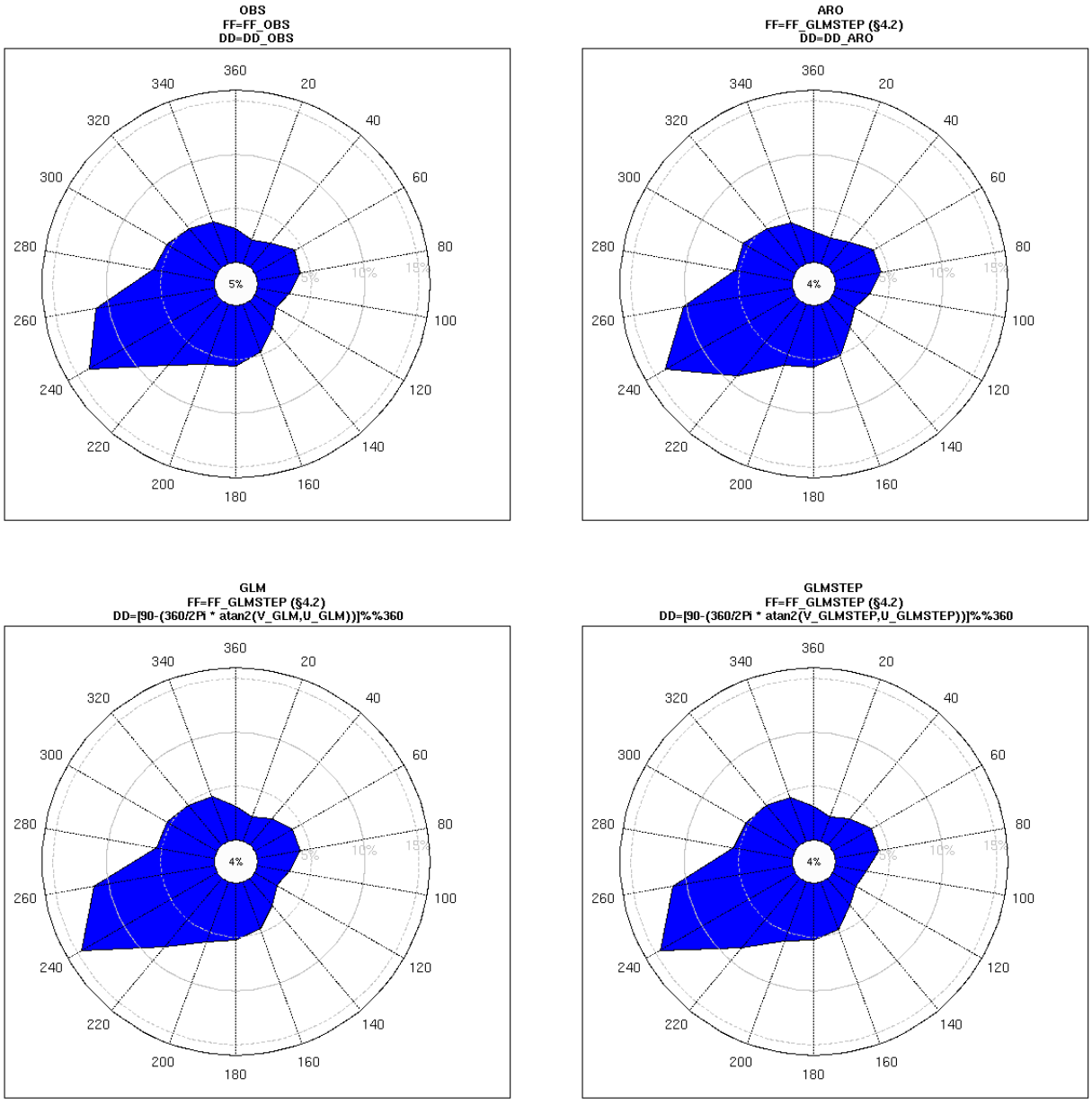


Illustration 18: Roses des vents après reconstitution de la direction DD à l'aide des composantes U et V sur l'échantillon test, toutes forces de vents (> 3 m/s) confondues. De gauche à droite et de haut en bas : observation (FF et DD OBS), BRUT (FF issue du l'extension (§4.2) et DD AROME), GLM (FF issue de l'extension (§4,2) et DD reconstitué à partir de U.GLM et V.GLM) et GLMSTEP (FF issue de l'extension (§4,2) et DD reconstitué à partir de U.GLMSTEP et V.GLMSTEP).

Les roses des vents sont très proches. La reconstitution statistique a tendance à trop lisser le secteur 160° - 200° par rapport au modèle BRUT. Cependant, les modèles GLM et GLMSTEP améliorent légèrement le modèle brut autour du secteur 260°.

Les scores B95+ (tableau 6), et notamment le score C2 qui concerne la direction, nous permettent d’évaluer la pertinence des modèles statistiques pour reconstituer la direction du vent. Le modèle GLMSTEP améliore le score C2.

Score B95+	BRUT	GLM	GLMSTEP
DD (C2)	96,77	97,61	97,80
Corrélation circulaire (C4)	94,02	94,44	94,41

Tableau 6: Scores B95+ sur les roses des vents BRUT, GLM et GLMSTEP.

Afin de valider ce résultat, nous allons calculer les scores de bonne prévision (H), de fausse alerte (FA) et de Pierce (PSS) pour chacune des tranches de secteurs (18 tranches de 20°) utilisées dans le score C2.

Plus le score de Pierce, qui évalue la qualité d’un modèle, est proche de 1, meilleur est le modèle. Dans l’illustration 19, le modèle BRUT et la GLMSTEP ont des scores supérieurs à 0, c’est-à-dire que le taux de bonne prévisions (H) est supérieur à celui de fausse alerte (FA).

Ces deux scores étant proche, il est difficile de différencier le modèle BRUT de la GLMSTEP. Il en est de même pour les scores de fausse alerte. Cependant, le modèle BRUT possède un meilleur score de bonne prévision que la GLMSTEP.

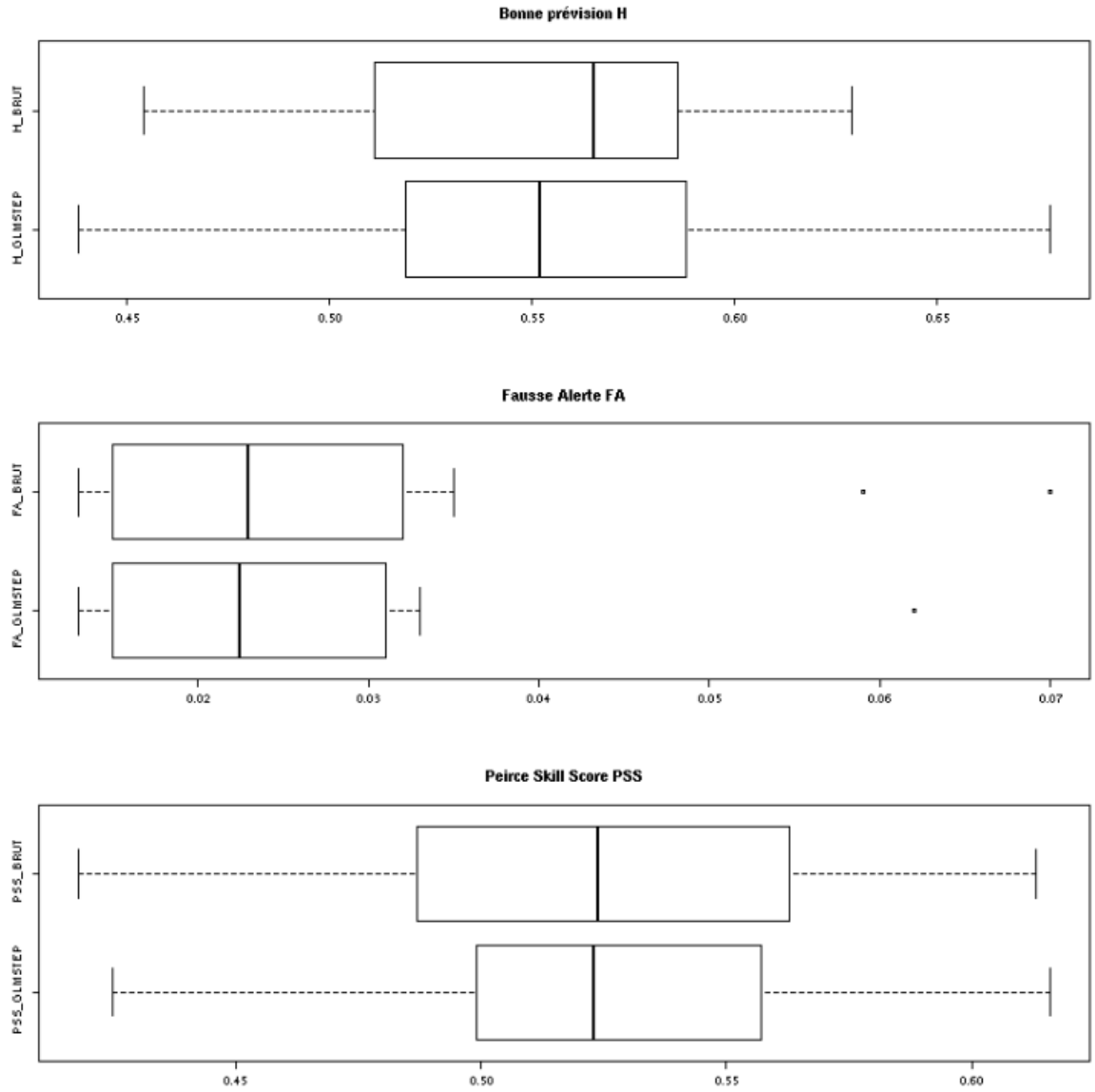


Illustration 19: Scores de bonne prévision (H), de fausse alerte (FA) et Pierce Skill Score (PSS). De haut en bas : H (H OBS/BRUT et H OBS/GLMSTEP), FA (FA OBS/BRUT et FA OBS/GLMSTEP) et PSS (PSS OBS/BRUT et PSS OBS/GLMSTEP).

En synthétisant ces différentes informations, nous **privilégions le modèle BRUT AROME** qui présente de très bons scores de qualité et de prévisions. Étant donné que les directions sont issues directement du modèle BRUT AROME, nous maintenons une cohérence temporelle dans les changements de directions.

4.4 Synthèse des choix de modélisation de DD et FF

A l'issue de l'étude sur la campagne de mesure, il a été retenu que **la force du vent sera étendue à l'aide du modèle statistique GLMSTEP** et que **la direction du vent sera issue du modèle AROME brut.**

L'illustration 20 présente les apports de cette modélisation.

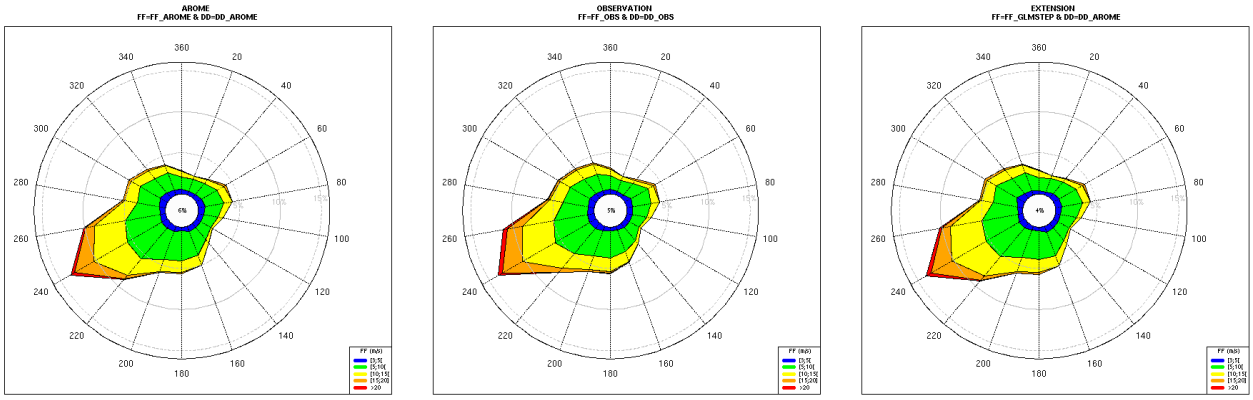


Illustration 20: Roses des vents sur la campagne de mesure. Gauche : Modèle AROME brut (FF AROME et DD AROME). Centre : Observation (FF OBS et DD OBS). Droite : Choix pour l'extension (FF GLMSTEP et DD AROME).

4.5 Extension de la série DD et FF à 100m

La série temporelle de direction (DD) et de force du vent (FF) à 100m est ensuite étendue sur la période du 01/01/2000 00H TU au 23/11/2016 23H TU, en appliquant le modèle statistique élaboré pour la force du vent (GLMSTEP) et le modèle brut AROME pour la direction du vent.

Un code qualité est appliqué pour chaque donnée :

- e = la donnée est estimée statistiquement (uniquement pour FF),
- b = la donnée est issue du modèle brut AROME,
- r = la donnée est reconstituée à partir des moyennes horaires des 3 jours précédents et des 3 jours suivants le manque.

Le code qualité de chaque composant (DD et FF) est fusionné de la façon suivante :

- si qualité de FF = e et qualité de DD = b → e
- si qualité de FF = e et qualité de DD = r → r
- si qualité de FF = b et qualité de DD = b → b
- si qualité de FF = b et qualité de DD = r → r
- si qualité de FF = r et qualité de DD = b → r
- si qualité de FF = r et qualité de DD = r → r

Il y a 147783 données estimées statistiquement, 155 données issues du modèle brut AROME et 190 données reconstituées sur la période du 01/01/2000 00H TU au 23/11/2016 23H TU.

La rose finale obtenue sur cette période est présentée dans l'illustration 21.

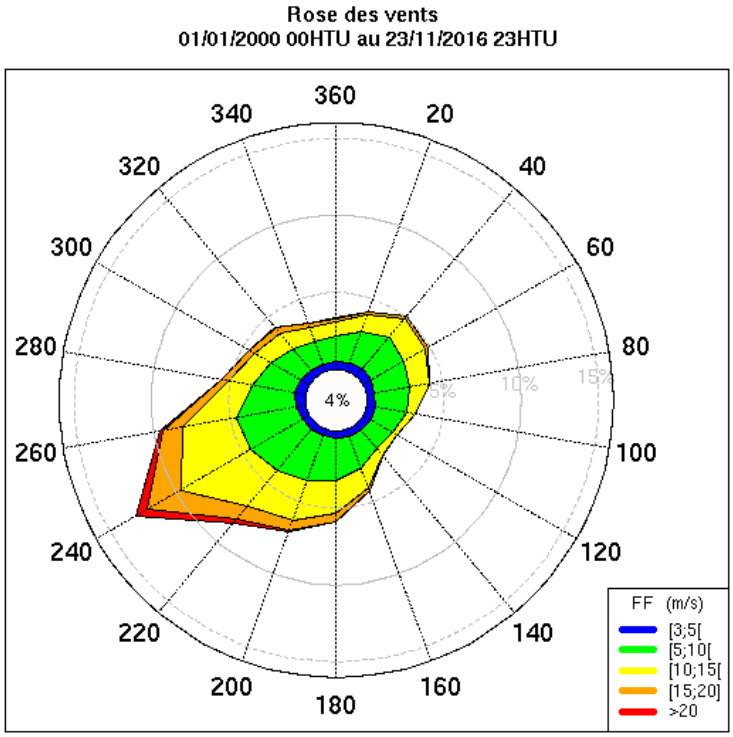


Illustration 21: Rose des vents - 01/01/2000 00HTU au 23/11/2016 23HTU

5 Limites d'utilisation des données reconstituées

Les données ont été reconstituées de sorte que les caractéristiques statistiques des distributions soient conformes à celle de l'observation. La cohérence statistique de la série a également été vérifié (cycle diurne, annuel, …) et la cohérence temporelle est maintenue au vu du coefficient de corrélation linéaire.

Nous attirons l'attention sur le fait que la série reconstituée reste une estimation statistique.

6 Livraison de la série temporelle d'observations horaires

La livraison des données consiste en un fichier .csv (le séparateur de colonne est « ; », le séparateur de décimale est « , ») :

- 17 ans de données par fichier (du 01/01/2000 00H TU au 23/11/2016 23TU),
- une ligne d'en-tête précisant le nom des colonnes,
- avec les colonnes suivantes :
  - DATE sous la forme AAAAMMJJHH (année, mois, jour, heure TU),
  - FF100 pour la force du vent à 100 m,
  - DD100 pour la direction du vent à 100m,
  - CODE, pour le code qualité associé à la donnée valant :
    - e si la donnée FF100 est estimée statistiquement et la donnée DD100 est issue du modèle brut,
    - b si la donnée est issue du modèle brut AROME,
    - r si la donnée est reconstituée à partir des moyennes horaires des 3 jours précédents et des 3 jours suivants le manque.

Nous avons appliqué les mêmes critères que pour l'observation, à savoir

- La force du vent à 100m est arrondie à la première décimale,
- La direction du vent à 100 est arrondie à l'entier (entre 0 et 359).

Le nom du fichier est le suivant :

- *extSerieLidarH100M.csv*

FIN DE DOCUMENT

Documents de référence antérieurs

	Intitulé	Référence	Date	Version
DR1	*Projet de parc éolien off-shore au large de Dunkerque*	Acquisition et suivi des mesures sur site durant un an	12/02/2018	V16